

From Lemmas to Dependencies: What Signals Drive Light Verbs Classification?

Sercan Karakaş
University of Chicago
skarakas@uchicago.edu

Yusuf Şimşek
Fırat University
ysimsek@firat.edu.tr

Abstract

Light verb constructions (LVCs) are a challenging class of verbal multiword expressions, especially in Turkish, where rich morphology and productive complex predicates create minimal contrasts between idiomatic predicate meanings and literal verb–argument uses. This paper asks what signals drive LVC classification by systematically restricting model inputs. Using UD-derived supervision, we compare lemma-driven baselines (lemma TF-IDF + Logistic Regression; BERTurk trained on lemma sequences), a grammar-only Logistic Regression over UD morphosyntax (UPOS/DEPREL/MORPH), and a full-input BERTurk baseline. We evaluate on a controlled diagnostic set with Random negatives, lexical controls (NLVC), and LVC positives, reporting split-wise performance to expose decision-boundary behavior. Results show that coarse morphosyntax alone is insufficient for robust LVC detection under controlled contrasts, while lexical identity supports LVC judgments but is sensitive to calibration and normalization choices. Overall, Our findings motivate targeted evaluation of Turkish MWEs and show that “lemma-only” is not a single, well-defined representation, but one that depends critically on how normalization is operationalized. We release the code and dataset to support reproducibility.¹

1 Introduction

Multiword expressions (MWEs), which can be described as word sequences that function as partially or fully lexicalized units, are a central property of human language and a persistent challenge for natural language processing, precisely because their meaning often departs from straightforward compositionality and must be learned as a unit (Sag et al., 2002; Odijk, 2013; Ramisch, 2015; Constant

et al., 2017; Barbu Mititelu et al., 2025). Recent work further shows that even strong contextualized models struggle to robustly capture idiomatic behavior: transformer-based models treat idiomatic expressions differently from simple lexemes, yet remain insensitive to graded differences in idiomaticity (Liu and Lareau, 2024), and VMWE identification continues to require specialized annotation schemes and evaluation pipelines even in high-quality UD resources (Markantonatou et al., 2025).

In comprehension and production, speakers are sensitive to distributional regularities of multiword sequences: frequent or strongly associated expressions can yield measurable processing advantages and facilitate learning, both in first and second language contexts (Arnon and Clark, 2011; Siyanova-Chanturia et al., 2011; Jiang and Siyanova-Chanturia, 2023; Siyanova-Chanturia and Sonbul, 2025). These behavioral findings motivate a computational perspective in which MWEs are not peripheral exceptions but core targets for models intended to approximate, support, or evaluate language use. Yet, MWEs systematically strain surface-based generalization: they can be semantically idiosyncratic, syntactically flexible, and highly variable under inflection, derivation, and word order alternations (Sag et al., 2002; Baldwin and Kim, 2010; Constant et al., 2017).

In NLP, MWE processing has traditionally split into *discovery* (mining candidate expressions from corpora) and *identification* (detecting and typing MWEs in context), with the latter typically formalized as supervised tagging or structured prediction (Ramisch, 2015; Constant et al., 2017). Community benchmarks such as the PARSEME shared tasks have sharpened the field’s focus on *verbal* MWEs (VMWEs), explicitly targeting discontinuity and morphosyntactic variability across languages (Savary et al., 2017, 2023). While transformer encoders have become strong default ar-

¹<https://github.com/ysufsimsek/lvc-lemma-vs-syntax>

chitectures for MWE identification, recent analyses emphasize that high token-level scores can mask brittle reliance on lexical templates or shallow heuristics, particularly for idiomatic and semi-lexicalized expressions (Premasiri and Ranasinghe, 2022; Miletić and Schulte im Walde, 2024). This gap between aggregate accuracy and robust semantic handling is especially salient for evaluation. Since many surface forms can realize the same content, automatic metrics are often hard to interpret and human-centered criteria are crucial for defining what constitutes model success (Zhou et al., 2022).

Turkish provides a demanding testbed for these issues. Its rich inflectional and derivational morphology multiplies surface variation, and productive complex predicate formation yields large families of verb–nominal predicates that range from transparent compositional combinations to strongly lexicalized idioms (Ofłazer, 1994; Ofłazer et al., 2004; Butt, 2010; Uçar, 2010). Among these, *light verb constructions* (LVCs) are particularly challenging: a nominal element contributes core predicational content while the verb is partially semantically bleached, and the entire expression behaves like a single predicate unit (Grimshaw and Mester, 1988; Sağ, 2015). Crucially, many Turkish light verbs also occur in fully literal transitive or ditransitive uses, creating hard minimal contrasts in which the same surface frame can express either a conventionalized predicate meaning or a literal event of transfer, perception, or action. This makes Turkish LVC detection a targeted probe of whether models encode the distinction between *predicate-level meaning* and *surface argument structure*. Although Turkish treebanks and MWE resources (including PARSEME) support supervised modeling, annotation conventions are not always exhaustive and practical performance can depend strongly on preprocessing quality (e.g., lemma and POS accuracy) (Eryiğit et al., 2011, 2015; Ozturk et al., 2022).

Lexical vs. morphosyntactic evidence for LVCs.

A central open issue in MWE modeling is where the evidence for MWE status resides: in lexical identity and selection (often well approximated by lemmas), or in morphosyntactic configuration (POS/morphology and dependency structure). This question mirrors a long-standing debate in psycholinguistics and acquisition about how multiword knowledge is represented: many accounts argue that speakers acquire and exploit *stored multiword units* (or strongly entrenched chunks) whose

availability is driven by frequency and association strength, yielding robust processing advantages for familiar sequences (Wray, 2002; Bannard and Matthews, 2008; Arnon and Snider, 2010). From this perspective, lemma-based representations are not merely an engineering simplification but a computational proxy for a lexicalist hypothesis: that MWE/LVC decisions can often be predicted from the identity of the participating lexemes and their conventionalized co-occurrence patterns. At the same time, MWEs, and especially verbal MWEs, also exhibit discontinuity, syntactic flexibility, and morphological variation, which motivates the use of morphosyntactic and dependency information both for candidate generation and for disambiguation in context (Savary et al., 2017; Simkó et al., 2017; Constant et al., 2017). Light verb constructions provide a particularly stringent test of the lexical–morphosyntactic division: because LVCs often share surface argument–structure frames with fully literal verb–argument configurations, broad structural cues can be insufficient unless a model also tracks predicate-level meaning and lexicalized selection restrictions. Consistently, prior work on LVC identification suggests that some cases are recoverable from syntactic regularities, while others require information beyond syntax alone (Cordeiro and Candito, 2019). Hence, reliable lemmatization and morphosyntactic preprocessing are themselves non-trivial, and resource-quality analyses for Turkish VMWE corpora emphasize that lemma/POS accuracy can be a major driver of identification performance (Ozturk et al., 2022). Crucially, recent work suggests that the same linguistic complexity also makes *tokenization* a first-order modeling choice rather than a neutral preprocessing step (Bayram et al., 2025a,b)

Moreover, experimental work on Turkish speech formulas and multiword expressions investigates whether familiar multiword strings show processing advantages over novel controls (Göymen and Aygüneş, 2020). Meanwhile, Turkish LVCs are highly productive and overlap structurally with literal transitive/ditransitive uses of the same light verbs, yielding hard minimal contrasts where the same syntactic frame can correspond to distinct predicate meanings (Grimshaw and Mester, 1988; Butt, 2010; Uçar, 2010). These properties make Turkish LVC detection an ideal setting for directly probing the relative contributions of lemma-level lexical content and morphosyntactic structure, including how comprehenders map surface form to

thematic/event interpretation (Özge et al., 2022).

Paper focus. Motivated by this, the present paper treats the lexical vs. morphosyntactic division not as a secondary engineering choice but as a primary scientific question for Turkish MWE modeling. We operationalize this question via controlled literal–idiomatic contrasts and by positioning models along a spectrum of available information, from lemma-restricted representations (emphasizing lexical selection and chunk-like knowledge) to morphosyntax-restricted representations (emphasizing grammatical configuration), with the goal of characterizing which signals are necessary and which are insufficient for robust LVC judgments. More broadly, this framing connects Turkish MWE identification to ongoing work on what kinds of linguistic structure models encode and how reliably such structure can be extracted or diagnosed with probing-style methods (Hewitt and Manning, 2019; Rogers et al., 2020; Katinskaia and Yangarber, 2024; Yoshida et al., 2024; Temesgen et al., 2025; Agarwal et al., 2025; Diego-Simón et al., 2025; Acevedo et al., 2026). In parallel, recent evaluation frameworks emphasize moving beyond single pooled scores toward more structured and contamination-resistant assessments that better separate genuine capability from memorization or artifacts (Cao et al., 2024).

Our diagnostic evaluation yields three main findings. First, coarse morphosyntactic information alone is largely sufficient to recognize broad negatives but is not sufficient for reliably identifying LVC positives under controlled literal–idiomatic contrasts. Second, lemma-level lexical identity provides substantially stronger evidence for LVC status, but lemma-based classifiers exhibit a calibration trade-off: improving performance on lexical controls can come at the cost of sharply increased false negatives on LVCs. Third, even for contextual encoders trained on lemma sequences, test-time behavior depends on how lemma normalization is instantiated, indicating that “lemma-only” is best understood as a family of representations with meaningful distribution shifts rather than a single input setting.

The remainder of the paper is organized as follows. Section 2 presents related work on MWEs/VMWEs and Turkish LVCs, with emphasis on lexical vs. morphosyntactic evidence and evaluation practices. Section 3 describes our UD-derived supervision, restricted-input representations, and

the construction of the controlled diagnostic set. Section 4 introduces the models and training setups. Section 5 details the experimental protocol and split-wise evaluation. Section 6 reports the results and discusses what they imply about the signals exploited by each model.

2 Related Work

As we have shown in the previous section, multiword expressions are not only a modeling nuisance but also a cognitive reality since a substantial portion of everyday language is produced and understood in partially prefabricated or strongly associated sequences (Wray, 2002; Conklin and Schmitt, 2012). Psycholinguistic and acquisition research has shown reliable processing advantages for frequent multiword units and sensitivity to phrase-level statistics, consistent with the view that speakers track distributional regularities beyond single words (Wray, 2002). This line of work motivates treating MWEs as *core* targets for NLP systems that aim to approximate or support human language use, rather than as peripheral exceptions to compositional grammar.

In NLP, MWE processing has traditionally separated *discovery* (mining candidate expressions via association measures or pattern heuristics) from *identification* (detecting and typing MWEs in context), with the latter typically formalized as supervised tagging or structured prediction (Ramisch, 2015; Constant et al., 2017). A persistent methodological question concerns *where* the evidence for MWE status resides: in lexical identity/selection (often approximated by lemmas and collocational strength) versus morphosyntactic configuration (POS/morphology and dependency structure). Practical systems therefore often combine lexical cues with morphosyntax and parsing, both to constrain candidate generation and to resolve discontinuity and variability in context (Constant et al., 2017; Savary et al., 2017; Simkó et al., 2017).

Recent work has largely converged on neural identification architectures, transformer encoders in particular, often benefiting from subword tokenization for robustness to inflectional variation (Bui and Savary, 2024; Ide et al., 2025; Hefetz et al., 2025; Rossini and van der Plas, 2026). At the same time, several recent analyses and surveys emphasize that strong aggregate scores can mask brittle reliance on shallow templates, memorized lexical cues, or dataset-specific artifacts, especially for

idiomatic and semi-lexicalized expressions where the core generalization is semantic (Premasiri and Ranasinghe, 2022; Milić and Schulte im Walde, 2024). This concern aligns with a broader trend in contemporary NLP and AI evaluation: controlled test suites and behavioral-style probes are increasingly used to localize *which* signals a model exploits, rather than only *how often* it is correct on pooled accuracy.

3 Methods

3.1 Training supervision from UD treebanks

To train supervised classification heads, we derived proxy supervision from nine Turkish Universal Dependencies (UD) treebanks released in CoNLL-U format: UD Turkish-GB, UD Turkish-IMST, UD Turkish-FrameNet, UD Turkish-BOUN, UD Turkish-PUD, UD Turkish-Penn, UD Turkish-Tourism, UD Turkish-Kenet, and UD Turkish-Atis (Universal Dependencies Consortium, 2025a,b,c,d,e,f,g,h,i).

These resources, comprising a total of 82,884 sentences, provide UD-standard tokenization, lemmas, UPOS tags, morphological features, and dependency relations, which we use both to derive weak supervision and to construct restricted-input representations.

To identify candidate LVC realizations from UD annotations, we exploited the dependency relations `compound:lvc` and `compound`. For treebanks that explicitly annotate LVCs with `compound:lvc`, we used these arcs directly as LVC candidates. For treebanks that do not contain `compound:lvc`, we followed an alternative procedure: (i) extract all `compound` dependencies, (ii) retain only noun-verb compound configurations (a nominal dependent linked to a verbal head), and (iii) manually review these candidates to remove non-LVC cases and to finalize a linguistically coherent set of target LVC patterns. Sentences containing validated LVC patterns were labeled as positive ([1]), and sentences without such patterns were labeled as negative ([0]).

To be even more precise, this pre-filtering step yielded 10,056 candidate LVC sentences. All candidates were then manually verified by two annotators using a binary label (LVC present vs. absent). During this screening, we identified 565 candidates that were incorrectly labeled as LVCs and removed them from the dataset. The resulting dataset contains 82,319 sentences, of which 9,491 contain at

least one LVC.

3.2 Restricted-input representations

It is worth recalling that a central methodological choice is to *restrict* what information the model can access in order to probe the relative contributions of lexical identity versus coarse morphosyntactic structure.

For lemma-restricted baselines, each sentence is converted to a sequence of lemma strings (preserving token order). This removes surface word-form variation while retaining lexical identity in normalized form and allowing contextual models to compose over the lemma sequence.

For grammar-restricted baselines, we remove lexical identity and keep only UD-style grammatical information: UPOS tags, dependency relations (DEPREL), and morphological features (FEATS). Each sentence is represented as a sentence-level count vector over these features (a bag of UPOS/DEPREL/MORPH features). Training uses the UD-derived supervision described above. At evaluation time, UD-style annotations for the diagnostic set are produced automatically, using Stanza (Qi et al., 2020), and features are extracted from these predicted annotations.

3.3 Task and evaluation set

Recall that we frame Turkish light verb construction detection as binary sentence classification: [1] indicates an LVC (including lexicalized/idiomatic noun+verb predicates under our labeling policy), and [0] indicates a literal verb-argument configuration. The controlled diagnostic evaluation set contains three balanced conditions ($n=49$ each; $N=147$): (i) RANDOM negatives: sentences that contain no LVC and are not constructed to reuse the specific light-verb lemmas targeted by the positive set; (ii) NLVC (*non-LVC*) lexical controls: in-domain sentences that reuse the same light-verb lemmas as the positives but enforce a *literal (non-idiomatic)* verb-argument reading; and (iii) LVC positives: sentences that contain an intended idiomatic light verb construction.

All evaluation items were manually written and then validated by three independent annotators for linguistic naturalness, pragmatic plausibility, and label agreement. This design prioritizes minimizing annotation noise and avoiding uncontrolled ambiguity over maximizing corpus breadth or domain diversity. Importantly, the diagnostic set was created specifically for this study rather than sampled

from an existing treebank or previously annotated corpus.

Although LVCs are often modeled via token-level identification (e.g., sequence tagging or arc labeling), we formulate the task as sentence-level *presence* detection to keep the comparison across restricted-input representations clean and assumption-light (no span, discontinuity, or decoding choices). This is especially relevant in Turkish because the same light verb can occur both in conventionalized LVC-like predicates and in fully literal uses (e.g., *vermek* ‘give’ in *öpücüük vermek* ‘give a kiss’ vs. literal transfer uses), creating hard minimal contrasts for Turkish (Özge et al., 2022) and other languages (Wittenberg and Piñango, 2011; Wittenberg et al., 2014; Wittenberg and Snedeker, 2014).

Despite the fact that $N = 147$ may appear modest, the dataset is purpose-built as a *controlled diagnostic* rather than an i.i.d. benchmark: it is balanced across conditions and lexically matched to isolate the LVC vs. literal distinction while minimizing confounds from topic, vocabulary, or annotation noise. This design choice follows a growing evaluation philosophy that prioritizes targeted stress tests and behavioral suites over aggregate i.i.d. accuracy, including checklist- and template-driven behavioral testing (Ribeiro et al., 2020; Naik et al., 2018; Yang et al., 2022; Zhao et al., 2024), contrast sets that probe local decision boundaries (Gardner et al., 2020), and dynamic/challenge-style benchmarks that surface systematic failure modes (Kiela et al., 2021). Related lines of work similarly argue that standard test splits can mask reliance on brittle heuristics or shortcuts (McCoy et al., 2019; Linzen, 2020; Geirhos et al., 2020). Finally, minimal-pair and diagnostic evaluation traditions show that carefully curated item sets can be disproportionately informative for pinpointing specific linguistic generalizations (Marvin and Linzen, 2018; Warstadt et al., 2020; Mueller et al., 2020; Mayne et al., 2025; Karakaş, 2026).

4 Models

4.1 Supervised BERTurk (full-input baseline)

We fine-tune BERTurk 32K cased and 128K cased (Schweter, 2020) by adding a task-specific binary classification head over the final-layer [CLS] representation. Training uses UD-derived proxy supervision with an 80/20 train/test split under stratified sampling to preserve the class distribution. For to-

kenization, we apply standard padding and truncation with a maximum sequence length of 64 tokens. To reduce overfitting, we set both hidden dropout and attention dropout to 0.2. Models are trained with learning rate 2×10^{-5} , batch size 32, and weight decay 0.01 for up to 10 epochs, monitoring validation loss with early stopping (patience 3) and selecting the checkpoint with the lowest validation loss. We report Accuracy, F1, and Loss. BERTurk 32K achieves 93.94% Accuracy, 0.7558 F1, and 0.1500 Loss; BERTurk 128K achieves 94.06% Accuracy, 0.7508 F1, and 0.1506 Loss, indicating comparable performance across vocabulary sizes. Notably, this baseline uses the original sentence forms as input (no lemmatization).

4.2 Lemma-only Logistic Regression (TF-IDF)

As a lightweight lexical baseline, we train Logistic Regression over TF-IDF features computed from lemma n -grams. Each sentence is replaced by its lemma sequence, and the resulting lemma text is lowercased prior to vectorization. We use unigram+bigram lemma features (`ngram_range=(1, 2)`) with `max_features=5000`. The classifier is L_2 -regularized Logistic Regression trained with `class_weight="balanced"` to mitigate class imbalance; we set `max_iter=1000` to ensure stable convergence. Supervised training follows the same 80/20 stratified split procedure. Evaluation is reported via overall accuracy and a standard classification report; model predictions for the diagnostic test set are exported for analysis.

4.3 Lemma-only BERTurk (restricted lexical input)

To isolate the effect of contextual composition under lemma normalization, we fine-tune BERTurk classifiers on inputs where each sentence is replaced by its lemma sequence, preserving token order and context while removing surface word-forms. This setting retains sequential context while explicitly removing surface-form variation due to inflection and orthography. We report lemma-only results for both BERTurk variants (32K and 128K) under the same training protocol as the full-input baseline (stratified 80/20 split, max length 64, padding/truncation, dropout 0.2, learning rate 2×10^{-5} , batch size 32, weight decay 0.01, early stopping with patience 3).

4.4 Grammar-only Logistic Regression (UPOS/DEPREL/MORPH)

To probe how far coarse morphosyntactic information goes without lexical identity, we train a Logistic Regression model using only UD-style grammatical annotations: UPOS tags, dependency relations, and morphological features. From the UD treebank training data, we extract (i) each UPOS tag, (ii) each dependency relation label, and (iii) each morphological key–value pair as a separate feature, and we construct a fixed feature inventory *only from the training portion* of the supervision. Each sentence is then represented as a sentence-level frequency vector counting occurrences of these UPOS/DEPREL/MORPH features. The classifier is trained as a binary Logistic Regression model with `class_weight="balanced"` and `max_iter=2000`.

5 Experiments

Our experiments aim to identify which information sources underlie Turkish LVC judgments by systematically restricting model inputs. We compare models along a continuum from lexical, lemma-based representations (lemma-only LR; lemma-only BERTurk) to grammar-only representations (LR over UPOS/DEPREL/MORPH), and include a full-input contextual baseline (BERTurk on surface sentences). This design enables a controlled assessment of normalized lexical identity, coarse morphosyntactic cues without lexical content and unrestricted contextual encoding.

5.1 Diagnostic evaluation set and split-wise reporting

We evaluate on a purpose-built diagnostic set of $N=147$ sentences, balanced across three conditions ($n=49$ each): RANDOM negatives, NLVC lexical controls that reuse the same light-verb lemmas in literal predicate–argument contexts, and LVC positives. This setup targets the key contrast—conventionalized predicate meaning vs. literal uses while controlling salient lexical cues via NLVC. Because we probe decision boundaries under minimal contrasts rather than maximize i.i.d. coverage, we report split-wise (Random/NLVC/LVC) results in addition to pooled scores; pooled accuracy can mask systematic misses of LVC under a conservative negative bias.

For the full-input BERTurk baseline, we evaluate each item in its original surface form (inflected

wordforms), mirroring standard VMWE/LVC identification.

Lemma-only BERTurk models are trained on lemmatized sentences (lemma sequences) and evaluated under two conditions: (i) *surface-test*, testing on original surface sentences (intentional train–test mismatch), and (ii) *lemma-test*, testing on a lemmatized version of the diagnostic set. This probes whether lemma-only models learn a stable lemma-level decision function or are sensitive to distribution shift at test time.

Grammar-only LR is trained on sentence-level count vectors over UPOS/DEPREL/MORPH features from UD training data. Since the diagnostic items are not from UD, we obtain UD-style annotations with Stanza and featurize them using the fixed training-only feature inventory; we additionally manually checked Stanza outputs for all 147 items to guard against systematic tagging/feature errors.

6 Results and Discussion

Lemma-only LR exhibits a clear calibration trade-off across runs: Runs 1–3 reflect different decision thresholds on the model’s posterior $p(y=1 | x)$ (Run 1 uses the default $\tau=0.5$, while Runs 2–3 use increasingly conservative thresholds τ^* chosen by sweeping $\tau \in [0, 1]$ on a held-out split using the precision–recall curve—e.g., selecting the τ^* that maximizes LVC F_1 and a higher- τ setting that prioritizes precision). As the classifier becomes more conservative, accuracy on negatives improves. Random remains near ceiling and NLVC can be pushed high while LVC accuracy drops markedly due to a rise in false negatives (Table 1). This pattern is consistent with a decision boundary that defaults to $[\emptyset]$ when lexical evidence is weak, yielding strong performance on negatives but limited sensitivity to LVC positives.

The grammar-only LR baseline makes the limitation of coarse morphosyntax especially explicit: despite near-perfect performance on Random and strong performance on NLVC, it collapses on LVC positives (Table 1). Given that (i) the diagnostic items are constructed to create minimal literal–idiomatic contrasts and (ii) Stanza analyses were manually checked for all 147 items, the most straightforward interpretation is that UPOS/DEPREL/MORPH frequency signatures alone are not sufficient to recover LVC status in these controlled contrasts. The model can

Model	Setting	Random	NLVC	LVC	Overall	FP	FN	Prec	Rec
Lemma-only LR									
	Run 1	98.0	75.5	49.0	74.1	13	25	64.9	49.0
	Run 2	98.0	87.8	32.7	72.8	7	33	69.6	32.7
	Run 3	100.0	95.9	18.4	71.4	2	40	81.8	18.4
BERTurk Full									
32k	Surface test	98.0	81.6	67.3	82.3	10	16	76.7	67.3
128k	Surface test	98.0	81.6	79.6	86.4	10	10	79.6	79.6
32k	Lemma test	100.0	79.6	63.3	81.0	10	18	75.6	63.3
128k	Lemma test	100.0	81.6	61.2	81.0	9	19	76.9	61.2
BERTurk (lemma-only)									
32K	Surface-test	98.0	87.8	67.3	84.4	7	16	82.5	67.3
128K	Surface-test	98.0	83.7	71.4	84.4	9	14	79.5	71.4
32K	Lemma-test	100.0	91.8	14.3	68.7	4	42	63.6	14.3
128K	Lemma-test	100.0	87.8	36.7	74.8	6	31	75.0	36.7
Grammar-only LR									
	–	100.0	91.8	10.2	67.3	4	44	55.6	10.2

Table 1: Restricted-input supervised baselines on the diagnostic set. Entries are success rates (%) for each condition ($n = 49$) and Overall ($N = 147$). FP/FN are pooled error counts where FP counts mistakes on negatives (Random+NLVC) and FN counts mistakes on LVC. Precision/Recall treat LVC as the positive class and pool Random+NLVC as negatives.

learn broad grammatical regularities that support negative decisions, but it largely fails to identify predicate-level conventionalization without lexical identity.

6.1 Lemma-only BERTurk: test-time input form matters

Lemma-only BERTurk shows a strong dependence on test-time input form (Table 1). When evaluated on surface sentences (surface-test), lemma-only BERTurk attains high accuracies across splits, including comparatively strong LVC performance. This suggests that even when supervision is provided on lemma sequences, the fine-tuned encoder can often generalize to inflected surface forms at inference time—a plausible outcome given subword tokenization overlap between lemmas and inflected variants.

However, when the same lemma-only models are evaluated on the lemmatized version of the diagnostic set (lemma-test), LVC accuracy drops sharply (especially for 32K; Table 1). This asymmetry suggests distribution shift between the lemma representation used for training and the lemma representation used at diagnostic evaluation. In our setup, lemma-only BERTurk is trained on UD-derived lemmas (treebank normalization conventions), whereas the lemmatized diagnostic set is produced by an automatic pipeline. Small mismatches in lemma inventory (e.g., normalization, casing, segmentation choices, named-entity handling, or token boundary differences) can substan-

tially change the subword sequences seen by the model, making lemma-test a different input distribution rather than a uniformly easier condition.

The 32K vs. 128K contrast is also suggestive: the 128K lemma-only model degrades less under lemma-test than the 32K model (Table 1; LVC recall 36.7% vs. 14.3%), consistent with the idea that a larger subword vocabulary can reduce token fragmentation (i.e., splitting lemmas into many smaller pieces) and thereby improve robustness to normalization- or lemma-form mismatches (Senrich et al., 2016; Devlin et al., 2019; Provilkov et al., 2020; Chizhov et al., 2024). Comparing BERTurk full and lemma-only under the same surface evaluation further reveals a vocabulary-dependent trade-off (Table 1): with 32K, lemma-only matches LVC recall while improving performance on negatives (higher NLVC accuracy and precision), yielding a slightly higher Overall score and fewer false positives. With 128K, the full model is clearly stronger on LVCs (higher recall, fewer false negatives) and improves Overall, consistent with surface-form supervision providing useful morpho-lexical cues when subword coverage is better. Notably, the 32K models show the opposite pattern: under surface-test, lemma-only slightly outperforms full overall despite identical LVC recall (Table 1), suggesting that at smaller subword vocabularies, incorporating surface-form detail can increase spurious matches on negatives without improving sensitivity to LVCs.

Evaluating the full BERTurk models on lemma-

normalized inputs further sharpens this interpretation (Table 1). Under lemma-test, both full models obtain the same Overall accuracy (81.0%), but LVC recall decreases relative to surface-test. Somewhat surprisingly, the 128K full model, despite being the stronger surface-form model, shows a much larger recall drop under lemmatization than the 32K full model (Table 1). This pattern suggests that the 128K model’s surface-form gains may rely more heavily on morpho-lexical cues present in inflected surface strings, which are attenuated or removed by lemma normalization. At the same time, full-on-lemma-test remains clearly stronger than lemma-only-on-lemma-test in LVC recall (Table 1), indicating that restricting training inputs to lemmas does not by itself yield robustness on LVCs in this diagnostic setting.

To sum up, the diagnostic results support three claims that motivate the paper. First, pure grammar signals (coarse UPOS/DEPREL/MORPH frequencies) can be sufficient to recognize negatives but are insufficient for robust LVC detection under controlled literal–idiomatic contrasts. Second, lemma-level lexical identity supports LVC judgments, but classical bag-of-lemmas models can be highly sensitive to calibration. Third, even when models are trained on lemma sequences, their behavior depends on how lemma normalization is instantiated at test time, which can be thought as a piece of evidence that lemma-only is not a single setting but a family of representations with potentially meaningful distribution shifts.

These patterns connect to several well-established threads in the MWE/VMWE literature. First, the weak performance of a grammar-only signal on LVC positives is consistent with the shared-task and survey view that VMWE identification typically requires combining lexical cues with morphosyntactic information: syntax helps with candidate generation and handling discontinuity/variability, but is often insufficient to separate idiomatic vs. literal readings when they share similar structural frames (Constant et al., 2017; Savary et al., 2017; Simkó et al., 2017; Cordeiro and Candito, 2019). Second, the strong dependence of lemma-only behavior on the exact normalization pipeline aligns with work highlighting that VMWE performance can be heavily shaped by preprocessing quality and annotation/representation choices, including Turkish-focused analyses that explicitly examine corpus/preprocessing effects for Turkish VMWEs

(Ozturk et al., 2022). Finally, the divergence between surface-test and lemma-test behavior reinforces a broader evaluation lesson in contemporary NLP: aggregate scores can obscure what cues models actually exploit, motivating controlled contrast sets and probing-style analyses that target decision boundaries and representation content (Hewitt and Manning, 2019; Gardner et al., 2020; Rogers et al., 2020).

7 Conclusion

This paper probed which signals support Turkish light verb construction (LVC) detection under controlled literal–idiomatic contrasts by restricting model inputs to lemmas, to grammar-only morphosyntax, or to full surface sentences. The results indicate that coarse UPOS/DEPREL/MORPH summaries can support negative decisions but are not sufficient for reliably recognizing LVC positives in this diagnostic setting, whereas lemma-level lexical information is more directly aligned with the predicate-level distinction the task requires. Results with BERTurk further underscore a vocabulary-dependent trade-off: lemma-only models can reduce false positives on negatives under surface evaluation, while full surface supervision can substantially improve LVC sensitivity when subword coverage is strong (Table 1). However, evaluating full models on lemma-normalized inputs reveals a non-trivial distribution shift, suggesting that high surface-form performance can rely on morpho-lexical cues that are attenuated by lemmatization.

At the same time, lemma-based modeling is not monolithic: bag-of-lemmas classifiers are sensitive to calibration, and lemma-only contextual encoders can be highly sensitive to how lemma normalization is operationalized at test time, revealing distribution-shift effects. These findings motivate split-wise reporting (Random/NLVC/LVC) as a default for LVC evaluation and position Turkish LVCs as a useful probe for separating lexicalized predicate meaning from surface argument structure in morphologically rich languages.

Limitations

Our supervision is UD-derived (compound:lvc and noun–verb compound), which is a practical proxy but may reflect treebank-specific conventions and incomplete coverage. The diagnostic set is small ($N=147$) by design, so results should

be read as decision-boundary evidence under controlled contrasts, not as in-the-wild performance. Grammar-only evaluation relies on Stanza parses (manually checked but not error free). Finally, lemma-only results can be sensitive to lemmatization/tokenization conventions, so “lemma-test” effects may partly reflect normalization choices rather than model competence alone.

References

- Santiago Acevedo, Alessandro Laio, and Marco Baroni. 2026. [Differential syntactic and semantic encoding in LLMs](#). *Preprint*, arXiv:2601.04765.
- Ananth Agarwal, Jasper Jian, Christopher D. Manning, and Shikhar Murty. 2025. [Mechanisms vs. outcomes: Probing for syntax fails to explain performance on targeted syntactic evaluations](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33725–33745, Suzhou, China. Association for Computational Linguistics.
- Inbal Arnon and Eve V. Clark. 2011. [Why brush your teeth is better than teeth: Children’s word production is facilitated in familiar sentence-frames](#). *Language Learning and Development*, 7(2):107–129.
- Inbal Arnon and Neal Snider. 2010. [More than words: Frequency effects for multi-word phrases](#). *Journal of Memory and Language*, 62(1):67–82.
- Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkha and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, 2 edition, chapter 12, pages 267–292. Chapman and Hall/CRC, Boca Raton, FL.
- Colin Bannard and Danielle Matthews. 2008. [Stored word sequences in language learning: The effect of familiarity on children’s repetition of four-word combinations](#). *Psychological Science*, 19(3):241–248.
- Verginica Barbu Mititelu, Maria Mitrofan, Mihaela Onofrei, Ioana Aioanei, and Liviu P. Dinu. 2025. [Survey on lexical resources focused on multiword expressions for the purposes of NLP](#). In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*. Association for Computational Linguistics.
- M. Ali Bayram, Ali Arda Fincan, Ahmet Semih Gümüŝ, Sercan Karakaŝ, Banu Diri, and Savaŝ Yıldırım. 2025a. [Tokenization standards and evaluation in natural language processing: A comparative analysis of large language models on Turkish](#). In *2025 33rd Signal Processing and Communications Applications Conference (SIU)*, pages 1–4, Istanbul, Turkey. IEEE. Conference held June 25–28, 2025. Preprint available as arXiv:2508.13058.
- M. Ali Bayram, Ali Arda Fincan, Ahmet Semih Gümüŝ, Sercan Karakaŝ, Banu Diri, Savaŝ Yıldırım, and Demircan Çelik. 2025b. [Tokens with meaning: A hybrid tokenization approach for NLP](#). *arXiv preprint arXiv:2508.14292*.
- Van-Tuan Bui and Agata Savary. 2024. [Cross-type French multiword expression identification with pre-trained masked language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4198–4204, Torino, Italia. ELRA and ICCL.
- Miriam Butt. 2010. The light verb jungle: Still hacking away. In Mengistu Amberber, Brett Baker, and Mark Harvey, editors, *Complex Predicates: Cross-linguistic Perspectives on Event Structure*. Cambridge University Press, Cambridge, UK.
- Boxi Cao, Mengjie Ren, Hongyu Lin, Xianpei Han, Feng Zhang, Junfeng Zhan, and Le Sun. 2024. [StructEval: Deepen and broaden large language model assessment via structured evaluation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5300–5318, Bangkok, Thailand. Association for Computational Linguistics.
- Pavel Chizhov, Catherine Arnett, Elizaveta Korotkova, and Ivan P. Yamshchikov. 2024. [BPE gets picky: Efficient vocabulary refinement during tokenizer training](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16587–16604, Miami, Florida, USA. Association for Computational Linguistics.
- Kathy Conklin and Norbert Schmitt. 2012. [The processing of formulaic language](#). *Annual Review of Applied Linguistics*, 32:45–61.
- Mathieu Constant, Gülŝen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. [Multiword expression processing: A survey](#). *Computational Linguistics*, 43(4):837–892.
- Silvio Ricardo Cordeiro and Marie Candito. 2019. [Syntax-based identification of light-verb constructions](#). In *Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 97–104, Turku, Finland. Linköping University Electronic Press.
- 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.
- Pablo J. Diego-Simón, Emmanuel Chemla, Jean-Rémi King, and Yair Lakretz. 2025. [Probing syntax in large language models: Successes and remaining challenges](#). In *Conference on Language Modeling (COLM) 2025*. ArXiv:2508.03211.

- Gülşen Eryiğit, Kübra Adali, Dilara Torunoğlu-Selamet, Umut Sulubacak, and Tuğba Pamay. 2015. [Annotation and extraction of multiword expressions in Turkish treebanks](#). In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 70–76, Denver, Colorado. Association for Computational Linguistics.
- Gülşen Eryiğit, Tuğba Ilbay, and Ozan Arkan Can. 2011. [Multiword expressions in statistical dependency parsing](#). In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 45–55. Association for Computational Linguistics.
- Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, and 7 others. 2020. [Evaluating models’ local decision boundaries via contrast sets](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. [Shortcut learning in deep neural networks](#). *Nature Machine Intelligence*, 2:665–673.
- Dilek Göymen and Mehmet Aygüneş. 2020. [The processing of speech formulas on turkish: A masked priming study](#). *Dilbilim Araştırmaları Dergisi*, 31(2):207–230.
- Jane Grimshaw and Armin Mester. 1988. [Light verbs and \$\theta\$ -marking](#). *Linguistic Inquiry*, 19(2):205–232.
- Ofri Hefetz, Kai Golan Hashiloni, Alon Mannor, and Kfir Bar. 2025. [Not just a piece of cake: Cross-lingual fine-tuning for idiom identification](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2521–2537, Mumbai, India. The Asian Federation of Natural Language Processing and The 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.
- Yusuke Ide, Joshua Tanner, Adam Nohejl, Jacob Hoffman, Justin Vasselli, Hidetaka Kamigaito, and Taro Watanabe. 2025. [CoAM: Corpus of all-type multiword expressions](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27004–27021, Vienna, Austria. Association for Computational Linguistics.
- Shang Jiang and Anna Siyanova-Chanturia. 2023. [The processing of multiword expressions in L1 and L2 Chinese: Evidence from reaction times and eye movements](#). *The Modern Language Journal*, 107(2):565–605.
- Sercan Karakaş. 2026. [Clause-internal or clause-external? testing turkish reflexive binding in adapted versus chain-of-thought large language models](#). *Preprint*, arXiv:2602.00380.
- Anisia Katinskaia and Roman Yangarber. 2024. [Probing the category of verbal aspect in transformer language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3347–3366, Mexico City, Mexico. Association for Computational Linguistics.
- Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ring-shia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. [Dynabench: Rethinking benchmarking in NLP](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.
- Tal Linzen. 2020. [How can we accelerate progress towards human-like linguistic generalization?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.
- Li Liu and Francois Lareau. 2024. [Assessing BERT’s sensitivity to idiomaticity](#). In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 14–23, Torino, Italia. ELRA and ICCL.
- Stella Markantonatou, Vivian Stamou, Stavros Bompoulas, Katerina Anastasopoulou, Irianna Linardaki Vasileiadi, Konstantinos Diamantopoulos, Yannis Kazos, and Antonios Anastasopoulos. 2025. [VMWE identification with models trained on GUD \(a UDv2 treebank of standard modern Greek\)](#). In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 14–20, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.
- Rebecca Marvin and Tal Linzen. 2018. [Targeted syntactic evaluation of language models](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

- Harry Mayne, Ryan Othniel Kearns, Yushi Yang, Andrew M. Bean, Eoin D. Delaney, Chris Russell, and Adam Mahdi. 2025. [LLMs don't know their own decision boundaries: The unreliability of self-generated counterfactual explanations](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24172–24197, Suzhou, China. Association for Computational Linguistics.
- R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. [Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.
- Filip Miletić and Sabine Schulte im Walde. 2024. [Semantics of multiword expressions in transformer-based models: A survey](#). *Transactions of the Association for Computational Linguistics*, 12:593–612.
- Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. [Cross-linguistic syntactic evaluation of word prediction models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.
- Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. [Stress test evaluation for natural language inference](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Jan Odijk. 2013. [Identification and lexical representation of multiword expressions](#). In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch: Results by the STEVIN Programme*, pages 201–217. Springer, Berlin, Heidelberg.
- Kemal Oflazer. 1994. [Two-level description of Turkish morphology](#). *Literary and Linguistic Computing*, 9(2):137–148.
- Kemal Oflazer, Bilge Say, Dilek Hakkani-Tür, and Gökhan Tür. 2004. [Integrating morphology with multiword expression processing for Turkish](#). In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing (MWE 2004)*, pages 64–71. Association for Computational Linguistics.
- Duygu Özge, Gülten Ünal, and İsa Kerem Bayırlı. 2022. [Assigning meaning to light verbs in Turkish](#). *Dilbilim Araştırmaları Dergisi*, 33(1):1–27.
- Yagmur Ozturk, Najet Hadj Mohamed, Adam Lion-Bouton, and Agata Savary. 2022. [Enhancing the PARSEME Turkish corpus of verbal multiword expressions](#). In *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022*, pages 98–105, Marseille, France. European Language Resources Association.
- Damith Premasiri and Tharindu Ranasinghe. 2022. [BERT\(s\) to detect multiword expressions](#). arXiv preprint.
- Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. [BPE-dropout: Simple and effective subword regularization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A Python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.
- Carlos Ramisch. 2015. *Multiword Expressions Acquisition: A Generic and Open Framework*. Theory and Applications of Natural Language Processing. Springer, Cham.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. [Beyond accuracy: Behavioral testing of NLP models with CheckList](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.
- 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.
- Diego Rossini and Lonneke van der Plas. 2026. [Binary token-level classification with DeBERTa for all-type MWE identification: A lightweight approach with linguistic enhancement](#). Preprint, arXiv:2601.19360. Accepted at Findings of EACL 2026.
- Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. [Multiword expressions: A pain in the neck for NLP](#). In *Computational Linguistics and Intelligent Text Processing (CICLing 2002)*, volume 2276 of *Lecture Notes in Computer Science*, pages 1–15, Berlin, Heidelberg. Springer.
- Yağmur Sağ. 2015. [Complex predicate formation via adjunction to a head category: Evidence from light verb constructions in Turkish](#). Manuscript, Rutgers University. Accessed: 2025-11-30.
- Agata Savary, Silvio Cordeiro, Carlos Ramisch, Federico Sangati, Veronika Vincze, Behrang Qasem-Zadeh, Marie Candito, Fabienne Cap, Voula Giouli, Regina Stodden, Antoine Doucet, and 1 others. 2017. [The PARSEME shared task on automatic identification of verbal multiword expressions](#). In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. Association for Computational Linguistics.

- Agata Savary, Carlos Ramisch, Silvio Cordeiro, and 1 others. 2023. [The PARSEME corpus release 1.3](#). In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*. Association for Computational Linguistics.
- Stefan Schweter. 2020. [BERTurk: BERT models for Turkish](#). Software release (Zenodo). Model cards: <https://huggingface.co/dbmdz/bert-base-turkish-cased> and <https://huggingface.co/dbmdz/bert-base-turkish-128k-cased>.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Neural machine translation of rare words with subword units](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Katalin Ilona Simkó, Viktória Kovács, and Veronika Vincze. 2017. [USzeged: Identifying verbal multiword expressions with POS tagging and parsing techniques](#). In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 48–53, Valencia, Spain. Association for Computational Linguistics.
- Anna Siyanova-Chanturia, Kathy Conklin, and Walter J. B. van Heuven. 2011. [Seeing a phrase “time and again” matters: The role of phrasal frequency in the processing of multiword sequences](#). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(3):776–784.
- Anna Siyanova-Chanturia and Suhad Sonbul. 2025. [Multiword expressions: New frontiers and challenges](#). *Language Teaching*.
- Tsedeniya Kinfé Temesgen, Marion Di Marco, and Alexander Fraser. 2025. [Extracting linguistic information from large language models: Syntactic relations and derivational knowledge](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27198–27214, Suzhou, China. Association for Computational Linguistics.
- Aygül Uçar. 2010. [Light verb constructions in Turkish dictionaries: Are they submeanings of polysemous verbs?](#) *Mersin Üniversitesi Dil ve Edebiyat Dergisi*, 7(1):1–17.
- Universal Dependencies Consortium. 2025a. [UD turkish-Atis](#). Universal Dependencies treebank hub page. Accessed: 2025-12-16.
- Universal Dependencies Consortium. 2025b. [UD turkish-BOUN](#). Universal Dependencies treebank hub page. Accessed: 2025-12-16.
- Universal Dependencies Consortium. 2025c. [UD turkish-FrameNet](#). Universal Dependencies treebank hub page. Accessed: 2025-12-16.
- Universal Dependencies Consortium. 2025d. [UD turkish-GB](#). Universal Dependencies treebank hub page. Accessed: 2025-12-16.
- Universal Dependencies Consortium. 2025e. [UD turkish-IMST](#). Universal Dependencies treebank hub page. Accessed: 2025-12-16.
- Universal Dependencies Consortium. 2025f. [UD turkish-Kenet](#). Universal Dependencies treebank hub page. Accessed: 2025-12-16.
- Universal Dependencies Consortium. 2025g. [UD turkish-Penn](#). Universal Dependencies treebank hub page. Accessed: 2025-12-16.
- Universal Dependencies Consortium. 2025h. [UD turkish-PUD](#). Universal Dependencies treebank hub page. Accessed: 2025-12-16.
- Universal Dependencies Consortium. 2025i. [UD turkish-Tourism](#). Universal Dependencies treebank hub page. Accessed: 2025-12-16.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohanney, 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.
- Eva Wittenberg, Martin Paczynski, Heike Wiese, Ray Jackendoff, and Gina R. Kuperberg. 2014. [The difference between “giving a rose” and “giving a kiss”: Sustained neural activity to the light verb construction](#). *Journal of Memory and Language*, 73:31–42.
- Eva Wittenberg and Maria Mercedes Piñango. 2011. [Processing light verb constructions](#). *The Mental Lexicon*, 6(3):393–413.
- Eva Wittenberg and Jesse Snedeker. 2014. [It takes two to kiss, but does it take three to give a kiss? categorization based on thematic roles](#). *Language, Cognition and Neuroscience*, 29(5):635–641.
- Alison Wray. 2002. *Formulaic Language and the Lexicon*. Cambridge University Press, Cambridge, UK.
- Guanqun Yang, Mirazul Haque, Qiaochu Song, Wei Yang, and Xueqing Liu. 2022. [TestAug: A framework for augmenting capability-based NLP tests](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3480–3495, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Ryo Yoshida, Taiga Someya, and Yohei Oseki. 2024. [Tree-planted transformers: Unidirectional transformer language models with implicit syntactic supervision](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5120–5134, Bangkok, Thailand. Association for Computational Linguistics.

Raoyuan Zhao, Abdullatif Köksal, Yihong Liu, Leonie Weissweiler, Anna Korhonen, and Hinrich Schütze. 2024. [SynthEval: Hybrid behavioral testing of NLP models with synthetic CheckLists](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7017–7034, Miami, Florida, USA. Association for Computational Linguistics.

Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022. [Deconstructing NLG evaluation: Evaluation practices, assumptions, and their implications](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States. Association for Computational Linguistics.