Modelling the Morphology of Verbal Paradigms: A Case Study in the Tokenization of Turkish and Hebrew

Giuseppe Samo
Idiap Research Institute
giuseppe.samo@idiap.ch

Paola Merlo
Idiap Research Institute
University of Geneva
paola.merlo@idiap.ch

Abstract

We investigate how transformer models represent complex verb paradigms in Turkish and Modern Hebrew, concentrating on how tokenization strategies shape this ability. Using the Blackbird Language Matrices task on natural data, we show that for Turkish –with its transparent morphological markers– both monolingual and multilingual models succeed, either when tokenization is atomic or when it breaks words into small subword units. For Hebrew, instead, monolingual and multilingual models diverge. A multilingual model using character-level tokenization fails to capture the language non-concatenative morphology, but a monolingual model with morpheme-aware segmentation performs well. Performance improves on more synthetic datasets, in all models.

1 Introduction

While language models excel at capturing distributional information at the token and sentence level (Warstadt et al., 2019; Linzen and Baroni, 2021; Gautam et al., 2024), their ability to generalize over paradigmatic phenomena, such as verb alternations (Levin, 1993; Kastner, 2019) and systematic patterns of verbal inflection remains less well understood (Yi et al., 2022; Proietti et al., 2022; Samo et al., 2023). The tokenisation step is an important aspect in understanding how a model treats verb alternations and paradigms in general, as tokenization shapes the internal representations in language models.

Paradigms capture the relational and systematic nature of linguistic elements, making variation meaningful and predictable within a broader structural framework (Setzke, 2024; Bobaljik, 2015). Consider a simple case: in many languages, causative verbs (verbs whose meaning implies that an actor caused the event described by the main verb) typically exhibit two voices, a transitive (T) and an intransitive (I) alternant.

Languages differ in the way they encode the voices of a paradigm (Haspelmath et al., 2014; Samardžić and Merlo, 2018). In English, neither alternant is morphologically marked (e.g., *the chef melts*$_T$ *the butter* vs. *the butter melts*$_I$), while in languages such as Italian only the intransitive form is morphologically marked (*scioglie*$_T$ vs. ***si** scioglie*$_I$ 'melts'). Conversely, in languages like Mongolian, the transitive form bears overt marking (*xajl-**uul**-ax*$_T$ vs. *xajl-ax*$_I$ 'melts'). In Japanese, both alternants are morphologically marked (*atum-**eru***$_T$ vs. *atum-**aru***$_I$ 'gather'). These differences are reflected in the internal model's representations, as the morphological marking affects the tokens and, consequently, the models' internal representations.

Morphological paradigms can be considerably complex, featuring larger inventories of voices—such as passive forms—and, consequently, a greater number of morphological markers. In this respect, Turkish provides a clear example of a system with transparent inflectional morphology, realized as a set of allomorphs attached to the verbal root in the form of affixes (Oflazer, 1993; Kornfilt, 1997; Göksel and Kerslake, 2005; Key, 2013).

Modern Hebrew exhibits a similarly complex verbal paradigm, but one characterized by a different type of morphological organization. Its well-studied patterns, known as *binyanim*, regulate how roots combine with morphological material to express a range of meanings, including causality (McCarthy, 1979; Arad, 2005; Tsarfaty, 2004). Although our data consist of Hebrew text without *niqqud*—the diacritic signs indicating vowels—some of the voices still display non-concatenative morphology.[1]

[1] Not all roots permit all templatic structures (Kastner, 2019). While roots typically convey a single overarching semantic field, the relationships between forms in a given paradigm are not always transparent (e.g., for the root PḲD, 'ordering'/'depositing'). This does not affect our investigation, which focuses on the morpho-syntactic aspects of the templates. Traditional grammars discuss seven binyanim, but

Form	Turkish	Hebrew	Gloss
Active	yazdı	כתב *katav* (PAAL)	wrote
Passive	yazıldı	נכתב *nixtav* (NIFAL)	was written
Causative	yazdırdı	הכתבי *hextiv* (HIFIL)	was dictated / made written
Causative & Passive	yazdırıldı	הכתב *huxtav* (HUFAL)	was dictated / was made written

Figure 1: Verbal paradigm voices under investigation and relative examples for the Turkish verb *yaz-* and the Hebrew root KTB (related to the act of writing). Hebrew binyanim are adapted from Kastner 2019, 574-575, in brackets the name of the binyanim.

The Turkish and Hebrew verbal paradigms under investigation are illustrated in Figure 1. In Turkish the active voice corresponds to the labile form of the inflected verb (e.g. person, number, tense features). The passive and causative voices are morphologically marked with affixes, while the causative-passive form combines both affixes. In Hebrew, the Paal binyan represents the basic, labile transitive form (cf. Turkish active). Nifal corresponds to the passive voice, often—but not always—marked by a prefix containing the character nun (Coffin and Bolozky, 2005, 71). Hifil and Hufal represent the causative active and passive voices, respectively; both can involve a prefix with the character *he*, but the active form is disambiguated by the presence of a *yod*—which marks a voiced palatal approximant and exemplifies non-concatenative morphology—alongside contextual cues.

Morphosyntactic cues are readily available to speakers (Fruchter and Marantz, 2015), but it remains unclear how they are captured by language models. Tokenization plays a key role in shaping the internal representations of language models (Hopton et al., 2025). However, standard substring tokenization could lead to different outcomes (Sennrich et al., 2016; Wu et al., 2016), depending on the granularity of the process and its coherence with the morphological distinctions. In one alternative, the verbal form might be tokenized in a linguistically-congruent way, clearly separating roots and inflections. In another option, it might be split into small sub-morphemes — sometimes as small as the grapheme level — fragmenting all the overt linguistic elements to the point where their relationship is no longer fully represented. In yet another possibility, the verbal form can be handled as one unit, keeping morphemes intact but blurring the distinction between root and inflection, and losing the compositional nature of morphological paradigms.

These alternative tokenization strategies determine what kind of morphological information is made available to the model directly and what morphological information needs instead to be induced in the hidden representations internal to the model. If a form is fragmented at the character level, the relation between root and pattern becomes opaque, as morphemes are broken into units too small to capture their function. If, instead, the morphemes appear in the same token, their internal structure is hidden. This motivates the use of a paradigm-level evaluation: by examining how models represent and process entire sets of related forms in complex settings, we can test whether morphological regularities survive tokenization choices and are encoded in sentence representations. The way words are tokenized affects their internal representations in sentences.

In this paper, we ask: Can current language models capture morphologically complex alternations in verbal paradigms in their internal representations, and how does tokenization affect their ability to represent these regularities?

To answer this question, we create structured datasets consisting of natural data (extracted from large-scale corpora) and synthetic data for a task appropriate for paradigms, the Blackbird Language Matrices (BLM) task (Merlo, 2023b), which is discussed in detail in Section 2. We evaluate the representations generated by transformer models using this task. The BLM task is paradigm-based and has been shown to be challenging, aiming to capture core morphosyntactic and semantic abilities of language models (Nissim et al., 2025).

2 The task

Blackbird Language Matrices (BLMs) are linguistic puzzles that implicitly describe paradigmatic linguistic systems (Merlo, 2023b; Merlo et al., 2022; Merlo, 2023a; An et al., 2023; Samo et al., 2023; Nastase et al., 2024a,b; Jiang et al., 2024). The task consists of a multiple-choice selecting the sentence that satisfies an underlying linguistic rule within a template. It has two components: (i) a context set of sentences that implicitly provides the information necessary to complete the linguistic paradigm, and (ii) an answer set of minimally differing con-

in this paper we focus on four, following Kastner (2019).

trastive sentences, where only one—the missing element in the pattern—is correct.

By analyzing sentence continuations that follow specific syntactic or semantic patterns, BLMs serve as an investigative tool for identifying systematic morphological and syntactic regularities. They provide an informative setup for studying how internal representations encode knowledge of linguistic paradigms, making them suitable for complex cases, such as the Turkish regularly compositional strings, and the Hebrew binyanim system, where verbal alternations are not easily distinguishable. To learn these alternations, the model must observe all variants and capture the relations among them, representing the compositional (Oflazer, 1993) or templatic structure of the system (Bobaljik, 2015). Templates are then instantiated creating curated datasets, and the task is then performed on sentence embeddings to test how linguistic information is encoded in the internal representations of language models. An example of a BLM template and its instantiation in Turkish and in Hebrew is provided in Figure 2. Details on the template and data instantiation are given in Section 3.

Previous BLM studies on concatenative languages have examined agreement in Romance and English (An et al., 2023; Nastase et al., 2024a) and verb alternations such as the English *spray/load* alternation (Samo et al., 2023) or the English and Italian causatives and object-drop phenomena (Nastase et al., 2024b). These studies show that representations rely heavily on morphological cues, yielding excellent task performance. However, they also depend on superficial, character-level signals, which hinders correct transfer across languages (Nastase et al., 2024b). Character-level signals are integral to phenomena such as the prototypical endings in agreement inflection, the restricted use of prepositions in *spray–load* constructions, and unique morphological markers like *si* in the intransitive form of Italian causatives.

3 Data and Models

In this section, we describe our structured, curated datasets, and its construction.[2] We also introduce the language models under investigation and their tokenization strategies.

3.1 BLM template

Each BLM template is composed of a context sentence set and an answer set. The context comprises three complete pairs of sentences with verbs inflected for a verbal form. The fourth pair is incomplete: one sentence illustrates the remaining form and the task consists in guessing the missing sentence. The answer set is composed of four sentences, each one illustrating a different form. Figure 2 shows an instantiation of the BLM template.[3]

3.2 Instantiation

Our dataset was created with natural occurring sentences extracted from treebanks annotated under the schema of the Universal Dependencies (UD; Nivre 2015; De Marneffe et al. 2021). We retrieved the sentences using *grew.match.fr*, by querying a simple variable X with the relevant annotation. The data for Turkish are collected from news and non-fiction sources (Penn v. 2.16[4]; 183,555 tokens, 16,396 trees) and grammar and dictionary examples (Kenet v. 2.16[5]; 178,658 tokens, 18,687 trees). The query collects sentences where the main verb is annotated with the VOICE parameter.[6] The Modern Hebrew data were extracted from two treebanks of Hebrew containing respectively news (HBT v.2.15, Tsarfaty 2013; McDonald et al. 2013; 114,648 tokens, 6,143 trees) and encyclopaedic entries (IAHLTWiki v. 2.15, henceforth IW; Zeldes et al. 2022; 103,395 tokens; 5,039 trees). The query collects sentences where the main verb is annotated with relevant the morphosyntactic property HEBBINYAN for Hebrew.[7]

These BLM-datasets contain the most complex type of lexical variation for the instantiation of

[2] Data is available at https://www.idiap.ch/en/scientific-research/data/blm-caush and https://www.idiap.ch/en/scientific-research/data/blm-caust.

[3] The presented examples are shown for illustrative purposes. At testing, to avoid potential order effects, the order of presentation of the context set pairs (apart from the last one) rotates over all possible orders of context for each possible correct answer.

[4] https://universaldependencies.org/treebanks/tr_penn/index.html

[5] https://universaldependencies.org/treebanks/tr_kenet/index.html

[6] The queries are as follows, ACT: *pattern* X [upos="VERB"]; X [VerbForm = "Fin"] *without* X [Voice = "CauPass" | "Cau" | "Pass" | "Rcp" | "Rfl"]; PASS: *pattern* X [upos="VERB"]; X [Voice = "Pass"]; X [VerbForm = "Fin"]; CAUS: *pattern* X [upos="VERB"]; X [Voice = "Cau"]; X [VerbForm = "Fin"]; CAUSPASS: *pattern* X [upos="VERB"]; X [Voice = "CausPass"]; X [VerbForm = "Fin"].

[7] The queries for Hebrew are as follows, ACT: *pattern* X [HebBinyan = "PAAL"]; PASS: *pattern* X [HebBinyan = "NIFAL"]; CAUS: *pattern* X [HebBinyan = "HIFIL"]; CAUSPASS: *pattern* X [HebBinyan = "HUFAL"].

TEMPLATE	TURKISH	MODERN HEBREW
	CONTEXT SET	
Act	*Şirketler kızgın bir şekilde kapasite ekledi .* Companies furiously added capacity (Penn, 15-0099.test)	*יציאתו של ליף גררה בתוך דקה את צימצום התוצאה ל-91 88.* ‘Lif's exit led.PAAL, within a minute, to narrowing the score to 91-88.’ (HTB, 5728)
Act	*Usta ağır çekim geldi.* The master has arrived in slow motion. (Kenet, 7702.train)	*אתי כיום עובדת כמוקדנית במוקד שירות לקוחות.* ‘Etti currently work.PAAL as a call center representative in a customer service center.’ (iahltwiki_hameila-babank-lemishar-76)
Pass	*1988'de madenlerde toplam 110000 ton bakır çıkarıldı .* In 1988, a total of 110,000 tons of copper were mined in the mines. (Penn, 15-0916.train)	*על כך נענשו שני הצדדים .* ‘For this, both sides were punished.’ (HTB, 5908)
Pass	*Değirmen taşları sert arkozdan yapılır .* Millstones are made of hard arkose. (Kenet, 0289.train)	*הגוש המזרחי ודרום אפריקה נמנעו בהצבעה .* ‘The Eastern Bloc and South Africa abstained in the vote.’ (iahltwiki_unrwa-14)
Caus	*Şirket 1967'den beri halı üretmektedir .* The company has been producing carpets since 1967. (Penn, 15-0153.train)	*אפשר כבר להכין את הכרטיסים .* ‘You can already prepare the tickets.’ (HTB, 5744)
Caus	*Bütün eski plaklar insanı hüzünlendirir .* All old records make you sad. (Kenet, 7911.train)	*בנוסף להשפעתו על המוח , משפיע האמפטמין גם על הגוף .* ‘In addition to its effect on the brain, amphetamine also affects the body.’ (iahltwiki_amphetamine-38)
CausPass	*Bu hisseler sonunda yeniden açıldı .* These shares eventually reopened. (Penn, 15-0010.test)	*שני ניצחונות בית צפויים הושגו בכפר בלום ובחולון .* ‘Two expected home victories were achieved in Kfar Blum and in Holon.’ (HTB, 5758)
ANSWER SET		
Act	*Sen gerçek hayattan bucak bucak kaçıyorsun .* You're running away from real life (Kenet, 7952.train)	*פריטוניטיס- דלקת בצפק , מוגדרת כעליה בתאי דם לבנים בנוזל הפריטונאלי .* ‘Peritonitis defines.PAAL as an increase in white blood cells in the peritoneal fluid.’ (iahltwiki_dialysis-39)
Pass	*İlk defa yeni usul bir rahleye oturtuldum .* For the first time, I was seated at a new style lectern. (Kenet, 5229.test)	*זיהומים הקשורים לקתטר* ‘Infections that associate.NIFAL with the catheter’ (iahltwiki_dialysis-44)
Caus	*Bu futbolcu antrenmanda göz doldurdu .* This football player impressed in training. (Kenet, 8210.train)	*על קיר זה פסיפס שנוצר בשנת 1990.* ‘On this wall, there is a mosaic created. HIFIL in 1990.’ (iahltwiki_holy-sepulchre-256)
CausPass	*Kaçakçılığın arkası alındı .* Smuggling was taken care of. (Kenet, 7783.train)	*ב-1985 הקליט איינשטיין עם שם טוב לוי את האלבום "תוצרת הארץ" .* ‘In 1985, Einstein recorded.HUFAL the album Totzeret Haaretz with Shem-Tov Levi’ (iahltwiki_arik-einstein-274)

Figure 2: BLM Template and instantiation in Turkish and Hebrew. The verb under investigation is underlined in the English translation. The indicated voice label is used only for error analysis, and not for training. The ID of the sentences refer to the dataset where the natural data are extracted as discussed in Section 3.

the BLM template—sentences in the same BLM sequence do not share a common, limited lexicon within one BLM instance (indicated as “type III” or “MaxLex” in other work on BLMs (An et al., 2023; Samo et al., 2023; Nastase et al., 2024a,b).

3.3 Models

To study the effect of tokenisation, we use sentence embeddings derived from both monolingual and multilingual models. We use the monolingual sentence embeddings of BERTurk for Turkish (*dbmdz/bert-base-turkish-cased*) and AlephBERT (*onlplab/alephbert-base*) for Hebrew. Following previous work on verb alternations (Yi et al., 2022; Samo et al., 2023; Nastase et al., 2024a), we also use the sentence embeddings of the multilingual model Electra (*google/electra-base-discriminator*, henceforth Multilingual/Multi).

We chose to work with transformer models rather than larger language or generative models for two reasons. First, transformers provide easy control over tokenization and embeddings. Second, using transformers allows us to obtain comparable, purely monolingual representations; by contrast, even large language models that can be described as monolingual do contain substantial amounts of English in their training data (Orlando et al., 2024).

4 Tokenisation

Tokenization plays a central role in shaping the internal representations of language models, making it essential to understand its impact on the modeling of linguistic phenomena. Most language models rely on sequential subword tokenization methods, such as Byte-Pair Encoding (Sennrich et al., 2016) or WordPiece (Wu et al., 2016), which have been proven effective for concatenative morphologies (Hopton et al., 2025).

As mentioned in the introduction, standard substring tokenization (Sennrich et al., 2016; Wu et al., 2016) can yield very different outcomes depending on its granularity and alignment with linguistic structure. A verbal form may be segmented in a morphologically meaningful way, split into very small sub-morphemic units that obscure functional relationships, or treated as a single token that preserves the surface form while hiding internal compositionality. We present our results in light of the tokenization strategies for the verbal form, which represent the primary focus of our investigation.

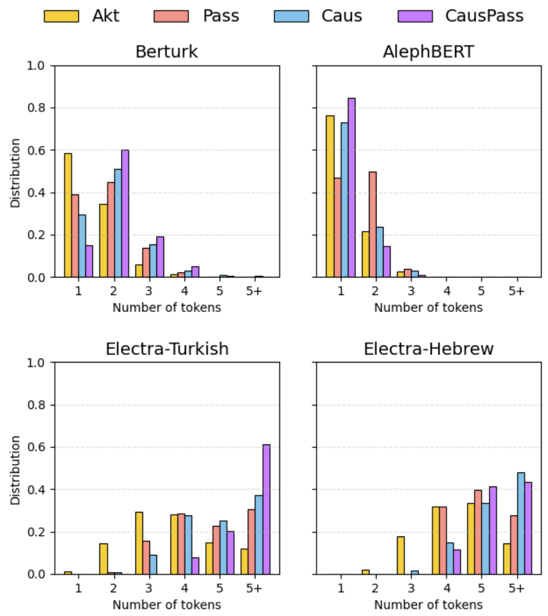

Figure 3: Number of tokens per voice forms across models and languages.

Voice	Berturk	Electra
Passive	##madı (91)	##ı (215)
	##du (80)	##r (68)
	##di (60)	##ld (54)
Causative	##dı (71)	##ı (1570)
	##di (49)	##r (455)
	##ıyor (44)	##yo (288)
Causative-Passive	##ildi (20)	##ı (1061)
	##ıldı (17)	##r (426)
	##ılıyor (8)	##yo (273)

Table 1: Turkish voices, top-3 most frequent tokens and their raw counts on the dataset (to be compared across models).

Figure 3 shows the number of tokens for each form. Monolingual models exhibit a similar pattern in both languages, but display a more atomic tokenization strategy for Hebrew verbal forms (an average of 1.329 tokens per form) than for Turkish (1.865 tokens per instance)[8]. In Turkish, however, the causative-passive form is the most segmented, showing more tokens on average than the other marked forms such as causative or passive.

For Hebrew, the multilingual Electra model exhibits a more character-based representation (5.136 tokens per verbal form, 1.001 characters per token). This difference, particularly for Hebrew script data, can be traced to the type of training data: Electra's fixed-size token vocabulary is dominated by frequent Latin-script languages, leaving only space for character-level representations of Hebrew (Muller et al., 2021; Ahia et al., 2023). On the other hand, the multilingual model in Turkish does not fully adopt a character-based tokenization (1.950 characters per token). Nevertheless, the number of tokens per instance increases with morphological complexity: 3.825 tokens for the labile (active) form, 4.911 and 5.222 for passive and causative forms, respectively, and 6.868 when the latter two markings are combined.

Finally, the linguistic informativeness of the tokens can be evaluated using the metrics in Table 1 for the three marked forms (Passive, Causative and Causative-Passive). Table 1 shows that the most frequent tokens in Turkish differ in quality between the monolingual and multilingual models for the marked forms. The table highlights not only differences in token quality but also in token length, with the multilingual model producing smaller subword units. In particular the most frequent tokens in the monolingual models are marker of tense and agreement (e.g. past tense *-dı* or progressive *-ıyor*). These results suggest that, in the more atomic monolingual model, the markers for passive and causative forms tend to remain attached to the root, while the multilingual model often splits them into separate subword tokens.

5 Experiments

We explore the behavior of a simple model through a series of experiments, reporting F1 scores and an error analysis.

5.1 Materials & Methods

Data Each dataset contains 8000 instances, split into 90:10 training:testing. We use disjoint sets of training and testing instances. In both datasets, the answers are equally distributed for each voice (1800 training: 200 testing). We ran four experiments for each dataset, isolating training and testing in answering each target voice.[9] Each run used 50 training epochs.[10]

System We used a Feed-Forward Neural Network (FFNN) as presented and discussed in the previous literature (Samo et al., 2023). By using a feed-forward neural network (FFNN), we test whether the semantic relations targeted by the task

[8] Table 2 in the Appendix provides details on tokenization for all datasets under investigation.

[9] A pilot study of three runs on each sub-dataset showed no differences in performance and errors across runs.

[10] All data for Hebrew were input as Hebrew alphabet characters without niqqud.

can be captured from the input representation. We aim to keep the system simple so that no other complex variables could explain the results. This is particularly appropriate in the context of this type of BLM task, since the two languages differ in how broad relational patterns are expressed: in Turkish, these relations are more transparently compositional, whereas in Hebrew, they are less so. For each sentence in the BLM, we use the averaged token embeddings. The FFNN takes in the stacked embeddings, uses a max-margin loss in training and selects the answer that has the highest cosine similarity to the output.

By applying a structure-agnostic architecture that operates over the entire input simultaneously, the FFNN allows us to test whether the semantic relations targeted by the task are recoverable directly from the geometry of the embedding space. This is particularly appropriate in the context of this type of BLM, where the challenge lies in identifying broad relational patterns rather than sequential dependencies.

5.2 Results

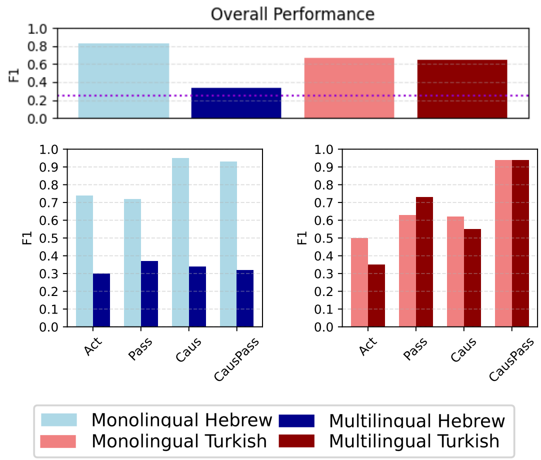

Figure 4: F1 for each voice as a correct answer across models. The dark violet dotted line in the upper panel indicates chance level.

Performance in terms of F1 scores is visualized in Figure 4. The overall performance is similar across models for Turkish. In Hebrew, however, the monolingual model (average F1 score: 0.835) significantly outperforms the multilingual model (0.333), with a large difference (Mann–Whitney U test: $U = 16.0$, $p = .029$, $r = 1.0$).

For Turkish, both the monolingual and multilingual models behave similarly: the active form is the most difficult to predict, while marked forms are easier—particularly the causative passive, which contains double marking. However, the linguistic quality of the tokenization discussed in section 3 does not introduce asymmetries. In Hebrew, the monolingual model performs better on the causative forms, but not consistently across all other marked forms such as passive. For the Hebrew multilingual model—which uses character-based tokenization—the binyanim in sentences are difficult to distinguish, resulting in low performance close to chance level.

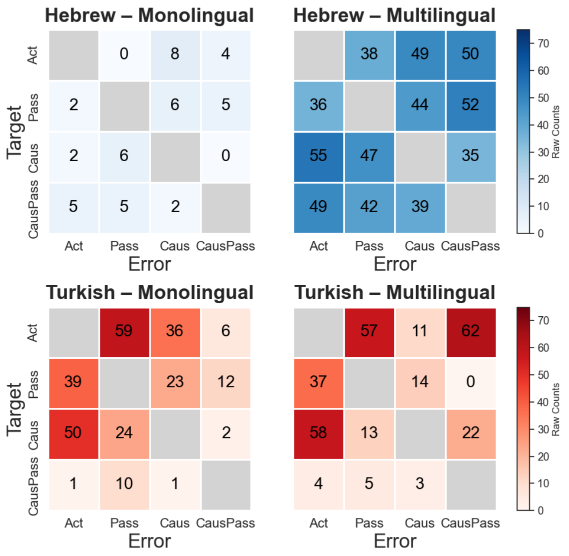

Figure 5: Confusion matrices of raw counts (test set n = 200)

Errors are visualized in the heatmap in Figure 5. We do not observe a consistently favoured answer across datasets and models except that in the Turkish monolingual, Passive is a prominent error when the target answer is Causative-Passive. For Hebrew the multilingual model shows a more distributed pattern of errors, consistent with chance-level performance. These results may indicate that granularity of tokenisation directly affects paradigm identification, at least for Hebrew. To better isolate the role of verbal paradigms, we create a synthetic dataset with reduced sentence complexity.

5.3 Analyzing the verbal paradigm

We created a second – more synthetic – dataset, which we label VERBONLY. This dataset contains only the verb corresponding to each sentence in the previous dataset, with no additional lexical content. The resulting sentences are grammatically correct, as both Hebrew and Turkish allow both subject and object drop (Vainikka and Levy, 1999; Erteschik-Shir et al., 2013; Meral, 2014).[11]

Specifically, this setup abstracts away the voices from both syntactic and lexical context, allowing us to focus exclusively on the verbal form and its morphological information. While this ultimately simplifies the task — since each sentence now contains only the verb, reducing its length and possibly noise — it allows for a more direct analysis of paradigms in the strict sense.

We run the same experiment on this synthetic dataset. Results are shown in Figure 6.

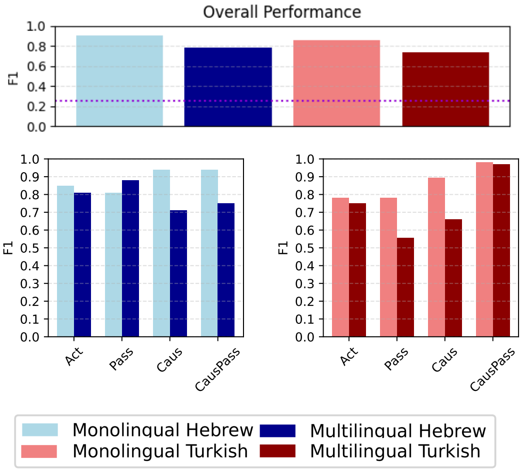

Figure 6: F1 for each binyan as a correct answer across models for the VERBONLY dataset. The dark violet dotted line indicates chance level.

In Turkish we observe an overall improvement of the performance, with the monolingual model performing slightly better than the multilingual in every voice. In Hebrew, the monolingual model is consistent with excellent results. The multilingual model also shows improved performance, approaching the monolingual model, although it still lags behind with respect to the causative forms.

As Figure 7 shows, in Turkish we observe a similar distribution of errors as the dataset containing full sentences. In Hebrew, however, one particularly informative type of error involves the confusion between Caus and CausPass, which share common morphological elements. Notably, in the multilingual VERBONLY model, which uses character-level tokenization, the most frequent error for target CausPass is Caus ($z = 4.31, p < .01$), and vice versa ($z = 5.35$, $p < .01$). This confusion suggests that

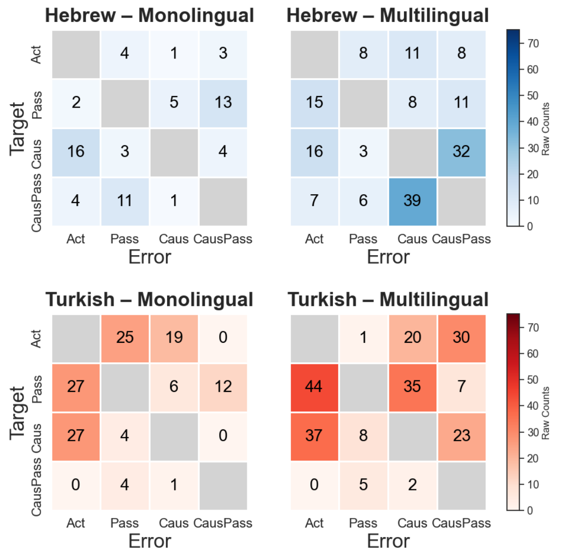

Figure 7: Confusion matrices of raw counts (test set $n = 200$) for the VERBONLY dataset.

the model's character-based representations capture surface-level morphological similarity and fail to fully distinguish deeper differences between the two binyanim.

5.4 Discussion

Our results provide a clear answer to our research question: transformer models can indeed capture morphologically complex alternations in their internal representations, but this capacity is highly dependent on how tokenization interacts with the specific morphological structure of a language. For Turkish—a language with transparent, concatenative morphology—both monolingual and multilingual models succeeded. The monolingual model employed a more atomic tokenization, often representing entire inflected word forms as single tokens, while the multilingual model used a more fragmented subword segmentation. Crucially, both strategies proved effective, indicating that for agglutinative systems, explicit surface forms—whether present in the atomic representation or in smaller segments—provide sufficient cues to infer paradigmatic relationships.

Conversely, the exact form of tokenization for Hebrew's morphology becomes decisive. The multilingual model based on character-level segmentation failed to capture the templatic binyanim system in natural sentences, performing near chance. Its overly fragmented tokenization may hinder the systematic co-dependence between root and pattern,

[11] The automatically retrieved inflected isolated forms in Hebrew may also contain affixes representing prepositions, determiners or complementizers (see also Shmidman and Rubinstein 2024).

rendering the paradigm opaque. The performance of the multilingual improves on simplified data, but it still lags behind the monolingual model.

These findings show that tokenization is not merely a preprocessing step, but possibly a linguistic filter that determines which morphological regularities are learnable. For the Turkish systems, both atomic and segmented tokenizations can be effective, but for Hebrew morphology, a representation that preserves or contains the integrity of the morphological form—whether atomic or appropriately segmented—is critical. The BLM task proves effective in diagnosing this interplay between tokenization strategy and morphological typology.

6 Related Work

The creation of paradigm-based datasets is useful for evaluating the generalization capacity of language models, for instance, across different morphological patterns (Batsuren et al., 2022; Nicolai et al., 2024; Warstadt et al., 2020). Verb alternations, especially in English, have been the object of recent investigation in LLMs showing excellent performance (Kann et al., 2019; Warstadt et al., 2019; Wilson et al., 2023). Yi et al. (2022) suggest that LLMs with contextual embeddings capture linguistic information about verb alternation classes at both the word and sentence levels in English. Also the semantic properties of the argument of verbs (agents, patients) have been tested with transformer models (Proietti et al., 2022).

The evaluation of LLMs' linguistic competence often relies on benchmark suites that make use of synthetic datasets. Synthetic datasets are constructed to probe specific grammatical phenomena in a controlled manner, frequently using minimal pairs or carefully designed paradigms (Warstadt et al., 2019, 2020). While automatic generation facilitates large-scale evaluation, it also raises concerns regarding distributional biases (Zhang and Pavlick, 2025; Nadas et al., 2025; Griffiths et al., 2024). In this paper, we begin our analysis by extracting data from large-scale naturalistic datasets, which may provide a more faithful basis for evaluating language models' representations of natural language (Jumelet et al., 2025).

The way tokenization is implemented (Rajaraman et al., 2024) also influences the accuracy of classification tasks in language identification and/or (neural) machine translation (Kanjirangat et al., 2023; Domingo et al., 2019). Approaches using neural encoders that operate directly on character sequences have been proposed and discussed (Clark et al., 2022). Hopton et al. (2025) demonstrated that subword tokenization can distinguish between function words (e.g., those indicating verbal constraints) and content words even in low-resource languages without annotated data.

Turkish has long been regarded as a key language for computational linguistics research on the interaction between morphology and tokenization, due to its highly agglutinative structure and productive inflectional system (Ataman et al., 2017; Ataman and Federico, 2018). Toraman et al. (2023) show that morphology-level tokenization for Turkish performs competitively with standard subword methods. Similarly, morphology-aware tokenization has been discussed in processing Semitic languages, proposing new tokenization algorithms (Goldman and Tsarfaty, 2022), such as linguistically informed extensions of BPE (Asgari et al., 2025). Gueta et al. (2023) integrate morphological knowledge directly into pretraining via specialized tokenization, showing gains on Hebrew across semantic and morphological benchmarks. According to Dang et al. (2024), both languages capture morphological knowledge effectively across various tokenization strategies, such as (sub)word-level and character-level approaches.

7 Conclusions

In this study, we examined how tokenization affects the ability of language models to represent complex verbal paradigms in Turkish and Hebrew. Our experiments show that tokenization granularity does interact with how internal sentence representations capture morphologically-complex alternations. Our results show that overall, monolingual models perform better than multilingual ones, indicating that performance is high when morphemes remain intact, whereas fragmentation can obscure systematic relations. Overall, these results underscore the importance of paradigm-level evaluation for understanding how models encode linguistic knowledge and highlight that tokenization strategy and language-specific morphology jointly shape internal representations. Future work should explore linguistically-informed tokenization schemes and extend these analyses to other morphologically-rich languages as well as to other linguistic phenomena to better understand the interaction between tokenization and linguistic knowledge in models.

Limitations

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

Ethics

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

Acknowledgments

We gratefully acknowledge the partial support of this work by the Swiss National Science Foundation, through grant SNF Advanced grant TMAG-1_209426 to PM. We thank Gökhan Özbulak and Ur Shlonsky for precious comments on the data.

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

Aixiu An, Chunyang Jiang, Maria A. Rodriguez, Vivi Nastase, and Paola Merlo. 2023. BLM-AgrF: A new French benchmark to investigate generalization of agreement in neural networks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1363–1374, Dubrovnik, Croatia.

Maya Arad. 2005. *Roots and Patterns: Hebrew Morpho-Syntax*. Springer, New York.

Ehsaneddin Asgari, Yassine El Kheir, and Mohammad Ali Sadraei Javaheri. 2025. Morphbpe: A morpho-aware tokenizer bridging linguistic complexity for efficient llm training across morphologies. *Preprint*, arXiv:2502.00894.

Duygu Ataman and Marcello Federico. 2018. An evaluation of two vocabulary reduction methods for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 97–110, Boston, MA. Association for Machine Translation in the Americas.

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically motivated vocabulary reduction for neural machine translation from turkish to english. *Preprint*, arXiv:1707.09879.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, and 76 others. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Jonathan David Bobaljik. 2015. Suppletion: Some theoretical implications. *Annu. Rev. Linguist.*, 1(1):1–18.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Edna Amir Coffin and Shmuel Bolozky. 2005. *A reference grammar of Modern Hebrew*. Cambridge University Press.

Thao Anh Dang, Limor Raviv, and Lukas Galke. 2024. Tokenization and morphology in multilingual language models: A comparative analysis of mt5 and byt5. *Preprint*, arXiv:2410.11627.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Miguel Domingo, Mercedes Garcıa-Martınez, Alexandre Helle, Francisco Casacuberta, and Manuel Herranz. 2019. How much does tokenization affect neural machine translation? *Preprint*, arXiv:1812.08621.

Nomi Erteschik-Shir, Lena Ibnbari, and Sharon Taube. 2013. Missing objects as topic drop. *Lingua*, 136:145–169.

Joseph Fruchter and Alec Marantz. 2015. Decomposition, lookup, and recombination: Meg evidence for the full decomposition model of complex visual word recognition. *Brain and Language*, 143:81–96.

Prakhar Gautam, Jitendra Singh Thakur, and Ashish Mishra. 2024. Subject–verb agreement error handling using rnn architectures. In *International Conference on Innovations in Computational Intelligence and Computer Vision*, pages 215–224. Springer.

Omer Goldman and Reut Tsarfaty. 2022. Morphology without borders: Clause-level morphology. *Transactions of the Association for Computational Linguistics*, 10:1455–1472.

Thomas L Griffiths, Jian-Qiao Zhu, Erin Grant, and R Thomas McCoy. 2024. Bayes in the age of intelligent machines. *Current Directions in Psychological Science*, 33(5):283–291.

Eylon Gueta, Omer Goldman, and Reut Tsarfaty. 2023. Explicit morphological knowledge improves pre-training of language models for hebrew. *Preprint*, arXiv:2311.00658.

Aslı Göksel and Celia Kerslake. 2005. *Turkish: A Comprehensive Grammar*. Routledge, London.

Martin Haspelmath, Andreea Calude, Michael Spagnol, Heiko Narrog, and Elif Bamyaci. 2014. Coding causal–noncausal verb alternations: A form–frequency correspondence explanation1. *Journal of linguistics*, 50(3):587–625.

Zachary William Hopton, Yves Scherrer, and Tanja Samardzic. 2025. Functional lexicon in subword tokenization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7839–7853, Albuquerque, New Mexico. Association for Computational Linguistics.

Chunyang Jiang, Giuseppe Samo, Vivi Nastase, and Paola Merlo. 2024. BLM-it - blackbird language matrices for Italian: A CALAMITA challenge. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 1135–1143, Pisa, Italy. CEUR Workshop Proceedings.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. MultiBLiMP 1.0: A Massively Multilingual Benchmark of Linguistic Minimal Pairs. *arXiv preprint arXiv:2504.02768*.

Vani Kanjirangat, Tanja Samardžić, Ljiljana Dolamic, and Fabio Rinaldi. 2023. Optimizing the size of subword vocabularies in dialect classification. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 14–30, Dubrovnik, Croatia. Association for Computational Linguistics.

Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R. Bowman. 2019. Verb argument structure alternations in word and sentence embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 287–297.

Itamar Kastner. 2019. Templatic morphology as an emergent property: Roots and functional heads in hebrew. *Natural Language & Linguistic Theory*, 37:571–619.

Gregory Key. 2013. *The morphosyntax of the Turkish caustive construction*. The University of Arizona.

Jaklin Kornfilt. 1997. *Turkish*. Routledge, London.

Beth Levin. 1993. *English Verb Classes and Alternations A Preliminary Investigation*. University of Chicago Press, Chicago and London.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(1):195–212.

John J. McCarthy. 1979. *Formal Problems in Semitic Phonology and Morphology*. Ph.D. thesis, Massachusetts Institute of Technology.

Ryan T McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B Hall, Slav Petrov, Hao Zhang, Oscar Täckström, and 1 others. 2013. Universal dependency annotation for multilingual parsing. In *Proc. of ACL*.

Hasan Mesut Meral. 2014. Silent objects and topic drop in turkish. *Dilbilim Araştırmaları A. Sumru Özsoy Armağanı*, 25:131–145.

Paola Merlo. 2023a. Blackbird language matrices (BLM), a new task for rule-like generalization in neural networks: Can Large Language Models pass the test? In *Findings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Paola Merlo. 2023b. Blackbird language matrices (blm), a new task for rule-like generalization in neural networks: Motivations and formal specifications. *Preprint*, arXiv:2306.11444.

Paola Merlo, Aixiu An, and Maria A. Rodriguez. 2022. Blackbird's language matrices (BLMs): a new benchmark to investigate disentangled generalisation in neural networks. *arXiv cs.CL.2205.10866*.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Mihai Nadas, Laura Diosan, and Andreea Tomescu. 2025. Synthetic data generation using large language models: Advances in text and code. *arXiv preprint arXiv:2503.14023*.

Vivi Nastase, Giuseppe Samo, Chunyang Jiang, and Paola Merlo. 2024a. Exploring italian sentence embeddings properties through multi-tasking. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (Clic-It 2024)*, pages 1–10.

Vivi Nastase, Giuseppe Samo, Chunyang Jiang, and Paola Merlo. 2024b. Exploring syntactic information in sentence embeddings through multilingual subject-verb agreement. In *Proceedings of the Tenth Italian*

Conference on Computational Linguistics (Clic-It 2024), pages 1–13.

Garrett Nicolai, Eleanor Chodroff, Frederic Mailhot, and Çağrı Çöltekin, editors. 2024. *Proceedings of the 21st SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, Mexico City, Mexico.

Malvina Nissim, Danilo Croce, Viviana Patti, Pierpaolo Basile, Giuseppe Attanasio, Elio Musacchio, Matteo Rinaldi, Federico Borazio, Maria Francis, Jacopo Gili, Daniel Scalena, Begoña Altuna, Ekhi Azurmendi, Valerio Basile, Luisa Bentivogli, Arianna Bisazza, Marianna Bolognesi, Dominique Brunato, Tommaso Caselli, and 62 others. 2025. Challenging the abilities of large language models in italian: a community initiative. *Preprint*, arXiv:2512.04759.

Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In *International conference on intelligent text processing and computational linguistics*, pages 3–16. Springer.

Kemal Oflazer. 1993. Two-level description of Turkish morphology. In *Sixth Conference of the European Chapter of the Association for Computational Linguistics*, Utrecht, The Netherlands. Association for Computational Linguistics.

Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. 2024. Minerva LLMs: The first family of large language models trained from scratch on Italian data. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 707–719, Pisa, Italy. CEUR Workshop Proceedings.

Mattia Proietti, Gianluca Lebani, and Alessandro Lenci. 2022. Does BERT recognize an agent? modeling Dowty's proto-roles with contextual embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4101–4112, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nived Rajaraman, Jiantao Jiao, and Kannan Ramchandran. 2024. Toward a theory of tokenization in llms. *Preprint*, arXiv:2404.08335.

Tanja Samardžić and Paola Merlo. 2018. The probability of external causation: An empirical account of crosslinguistic variation in lexical causatives. *Linguistics*, 56(5):895–938.

Giuseppe Samo, Vivi Nastase, Chunyang Jiang, and Paola Merlo. 2023. BLM-s/lE: A structured dataset of English spray-load verb alternations for testing generalization in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12276–12287, Singapore. Association for Computational Linguistics.

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.

Rafael Soto Setzke. 2024. Linguistic paradigms as cognitive entities: A domain-general approach. *Yearbook of the German Cognitive Linguistics Association*, 12(1):73–94.

Avi Shmidman and Aynat Rubinstein. 2024. Computational methods for the analysis of complementizer variability in language and literature: The case of Hebrew "she-" and "ki". In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 294–307, Miami, USA. Association for Computational Linguistics.

Cagri Toraman, Eyup Halit Yilmaz, Furkan Şahiṇuç, and Oguzhan Ozcelik. 2023. Impact of tokenization on language models: An analysis for turkish. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–21.

Reut Tsarfaty. 2004. 'binyanim ba'avir': An investigation of Aspect Semantics in Modern Hebrew. Master's thesis, Universiteit van Amsterdam, Institute for Logic Language and Computation, December.

Reut Tsarfaty. 2013. A unified morpho-syntactic scheme of stanford dependencies. In *Proc. of ACL*.

Anne Vainikka and Yonata Levy. 1999. Empty subjects in finnish and hebrew. *Natural Language & Linguistic Theory*, 17(3):613–671.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2877–2887, Hong Kong, China. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Michael Wilson, Jackson Petty, and Robert Frank. 2023. How abstract is linguistic generalization in large language models? experiments with argument structure. *Transactions of the Association for Computational Linguistics*, 11:1377–1395.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, and 1 others. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

David Yi, James Bruno, Jiayu Han, Peter Zukerman, and Shane Steinert-Threlkeld. 2022. Probing for understanding of English verb classes and alternations in large pre-trained language models. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 142–152, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Amir Zeldes, Nick Howell, Noam Ordan, and Yifat Ben Moshe. 2022. A second wave of UD Hebrew treebanking and cross-domain parsing. In *Proceedings of EMNLP 2022*, pages 4331–4344, Abu Dhabi, UAE.

Lingze Zhang and Ellie Pavlick. 2025. Does training on synthetic data make models less robust?

Appendix

		SENTENCES				VERBS				
VOICE	INST	CH	TOK	TOK/INST	CH/TOK	CH	TOK	TOK/INST	CH/TOK	1V-TOK
HEBREW - MONOLINGUAL										
Total	**6899**	**880210**	**215195**	**31.130**	**4.093**	**35140**	**8966**	**1.329**	**3.888**	**3746**
Act	1928	242839	60117	31.181	4.039	8652	2254	1.169	3.839	1155
Pass	1892	234444	57435	30.357	4.082	9516	2478	1.310	3.840	897
Caus	1972	262825	63653	32.278	4.129	10815	2495	1.265	4.335	1104
CausPass	1107	140102	33990	30.705	4.122	6157	1739	1.571	3.541	590
HEBREW - MULTILINGUAL										
Total	**6899**	**880210**	**711294**	**102.993**	**1.237**	**35140**	**35101**	**5.136**	**1.001**	**169**
Act	1928	242839	195663	101.485	1.241	8652	8630	4.476	1.003	49
Pass	1892	234444	189327	100.067	1.238	9516	9512	5.027	1.000	41
Caus	1972	262825	212857	107.940	1.235	10815	10802	5.478	1.001	40
CausPass	1107	140102	113447	102.481	1.235	6157	6157	5.562	1.000	39
TURKISH - MONOLINGUAL										
Total	**5112**	**387195**	**81522**	**15.984**	**4.784**	**46379**	**8952**	**1.865**	**5.391**	**3044**
Act	1841	139698	29663	16.112	4.710	14002	2803	1.523	4.995	1341
Pass	1879	138224	29006	15.437	4.765	17688	3377	1.797	5.238	870
Caus	1240	97437	20435	16.480	4.768	12661	2442	1.969	5.185	657
CausPass	152	11836	2418	15.908	4.895	2028	330	2.171	6.145	176
TURKISH - MULTILINGUAL										
Total	**5112**	**387195**	**162347**	**32.148**	**2.379**	**46379**	**23793**	**5.207**	**1.950**	**1847**
Act	1841	139698	58299	31.667	2.396	14002	7042	3.825	1.988	698
Pass	1879	138224	57846	30.786	2.390	17688	9228	4.911	1.917	480
Caus	1240	97437	41199	33.225	2.365	12661	6479	5.225	1.954	465
CausPass	152	11836	5003	32.914	2.366	2028	1044	6.868	1.943	204

Table 2: Comparison of AlephBERT, Berturk and ELECTRA tokenization of sentences and inflected verbs in terms of number of tokens (TOK), number of characters (CH), token per word (TOK/W), characters per token (CH/TOK) and one-token verbal forms (1V-TOK). The highlighted cells in grey for the tokenization of verbs refer to the increasing number of token pro verbal form.