The Degree of Language Diacriticity and Its Effect on Tasks

Adi Cohen, Yuval Pinter
Institute for Applied AI Research
Ben-Gurion University of the Negev
Beer Sheva, Israel
{adibc@post, uvp@cs}.bgu.ac.il

Abstract

Diacritics are orthographic marks that clarify pronunciation, distinguish similar words, or alter meaning. They play a central role in many writing systems, yet their impact on language technology has not been systematically quantified across scripts. While prior work has examined diacritics in individual languages, there's no cross-linguistic, data-driven framework for measuring the degree to which writing systems rely on them and how this affects downstream tasks. We propose a data-driven framework for quantifying diacritic complexity using corpus-level, information-theoretic metrics that capture the frequency, ambiguity, and structural diversity of character-diacritic combinations. We compute these metrics over 24 corpora in 15 languages, spanning both single- and multi-diacritic scripts. We then examine how diacritic complexity correlates with performance on the task of diacritics restoration, evaluating BERT- and RNN-based models. We find that across languages, higher diacritic complexity is strongly associated with lower restoration accuracy. In single-diacritic scripts, where character-diacritic combinations are more predictable, frequency-based and structural measures largely align. In multi-diacritic scripts, however, structural complexity exhibits the strongest association with performance, surpassing frequency-based measures. These findings show that measurable properties of diacritic usage influence the performance of diacritic restoration models, demonstrating that orthographic complexity is not only descriptive but functionally relevant for modeling.

Keywords: diacritics, writing systems, character-level models, character tagging

1. Introduction

Diacritics have been known to affect performance of language technology systems, but a thorough data-driven survey has yet to be done. Gorman and Pinter (2025) started to explore this issue, presenting anecdotal evidence and providing rules of thumb for NLP practitioners.

In this study, we consider the presence of diacritics a property of the writing system of a language and build a bottom-up framework for quantifying the degree to which diacritics play a role in corpora of a language, leading to a mechanism for examining the *degree* to which diacritics affect task performance, for a task expected to present such variation, namely diacritization itself.

Our approach considers "diacritifulness" of a langauge along a spectrum: for example, in Spanish vowels occasionally feature one kind of an accent mark (e.g., á), and one base consonant character (n) is shared by two phonemes (n and ñ). In Vietnamese, both vowel quality and tone are marked with co-inciding marks. In Hebrew, vowels are marked on phonetically preceding consonant characters and exist in large combinatorial distribution. Once this variation is accounted for quantitatively, the true effects of a language writing system on technology performance can be examined via tools like correlation.

Concretely, we propose the first-ever attempt to quantify the degree to which a language uses diacritics in its canonical writing system. Our approach is data-driven and information-theoretic: we calculate corpus-level metrics based on the distributions of diacritic marks, characters, and their combinations in each language. We assess the effect this property of "diacritifulness" has on the closest task to the subject matter: *diacritization* of undiacritized text, measured by word-level and character-level accuracy. Our results, identifying the correlation between a language's diacritifulness and success of various neural diacritization models, suggest that consideration of whether each of a script's characters admits one or more diacritic at a time, which we term single- vs. multi-diacritic systems, affects this relationship greatly, particularly for transformer-based diacritization models. We hope that our framework encourages more work on gauging written properties of languages and their effects on tasks, perhaps more downstream ones such as question answering and machine translation.[1]

2. Background

Diacritics are marks attached to base letters that encode additional phonological and/or grammatical information. They may indicate vowel quality or length, tone, gemination, or morphological distinctions. In some languages, like Vietnamese, diacritics are essential and form part of standard spelling.

[1] Our code, models and data are available at https://github.com/MeLeLBGU/Diacriticity.

Proceedings of the Third Workshop on Computation and Written Language (CAWL 2026) @LREC 2026, pages 1–8
12 May, 2026. ©ELRA Language Resources Association (ELRA), 2026

In other languages, such as Arabic and Hebrew, diacritics encode important phonological or grammatical distinctions but are often omitted in everyday writing. Writing systems that omit some or all diacritic marks are often referred to as defective writing systems, since the written text does not fully specify pronunciation or lexical identity.

The omission of diacritics introduces ambiguity. A single undiacritized form may correspond to multiple valid interpretations, requiring the inference of the intended meaning from context. This has been cited as one major reason that speech and language technologies for Arabic and Hebrew seemingly lag behind other, similarly-resourced languages (e.g., Tsarfaty et al., 2019).

In text-based systems, including machine translation, different diacritizations may have several valid lexical or grammatical meanings. Despite this, many NLP pipelines remove diacritics during preprocessing, especially in multilingual transformer-based models such as mBERT (Clark et al., 2022).

Diacritization, the task of restoring missing diacritics, has been extensively studied, particularly for Arabic and Hebrew. Early approaches relied on rules, lexicons or morphological analyzers (e.g., Darwish et al., 2017; Krstev et al., 2018). More recent work formulates diacritization as a sequence tagging or sequence-to-sequence problem using neural architectures such as RNNs and transformers (e.g., Belinkov and Glass, 2015; Shmidman et al., 2020; Gershuni and Pinter, 2022; Náplava et al., 2018; Náplava et al., 2021).

While these models perform well within individual languages, most work remains focused on language-specific or single writing system, such as Latin-based scripts. As a result, cross-script comparisons remain limited. There has also been relatively little systematic investigation of how orthographic properties themselves, such as diacritic frequency, structural diversity or ambiguity, affect restoration difficulty across languages.

Languages differ in how diacritics are used. In some scripts, each character can take only a single diacritic, while in others, base characters can carry multiple simultaneous marks. We refer to characters carrying exactly one diacritic as **single-diacritic**, and to those carrying two or more simultaneous marks as **multi-diacritic**. At the language level, we classify a writing system as multi-diacritic if it permits any character to bear multiple simultaneous diacritics; otherwise, we classify it as single-diacritic.

Beyond frequency and function, languages also vary in the predictability and diversity of their diacritic patterns. These differences suggest that diacritic usage varies along a spectrum of structural complexity. We propose quantitative corpus-level metrics to capture this variation and examine how it relates to neural diacritization performance across languages.

3. Metrics

Let C be the set of base characters of a language (e.g., the Latin letters a–z), and let D be the set of diacritic marks that may attach to them. We define a *rune* as a base character together with zero or more diacritics, so that both unmarked and diacritized characters are treated uniformly. For example, the Spanish character á can be viewed as the base letter a with an acute accent, while the Hebrew form בְּ consists of the base letter ב with two diacritic marks ◌ְ and ◌ּ. In both cases, we treat the base character together with its diacritics as a single rune. A diacritized string is then a sequence of runes $r_1, \ldots, r_n$, where each rune r_i consists of a base character $c_i \in C$ and a (possibly empty) subset of diacritics from D.

We follow previous work in defining a stripping function σ that removes diacritics and returns the corresponding base character sequence $c_1, \ldots, c_n$.

To capture the deeper properties of writing scripts and diacritics, we introduce a set of information-theoretic metrics that quantify the relationship between base characters and their diacritized realizations.

We present examples for our metrics on Hebrew and Spanish in Table 1.

Rune Surprisal. Rune surprisal (RS) quantifies the level of uncertainty associated with a diacritized version based on its base character. It is high when a character appears with multiple competing diacritic forms of similar probability, and low when a single form clearly dominates.

Formally, let r be a rune with base character $\sigma([r]) = c$, we define:

$$P(r \mid c) = \frac{\#(r)}{\sum_{r' : \text{base}(r')=c} \#(r')}.$$

The suprisal of a rune is:

$$RS(r) = -\log P(r \mid c).$$

Languages that show higher average rune surprisal when calculated over a corpus display greater ambiguity per character, which we expect to make diacritization more difficult.

Diacritic Token Surprisal. Diacritic token surprisal (DTS) breaks each rune into its individual diacritics marks and measures how unexpected those marks are given the base character. Rather than treating the full diacritized form as a single unit, this

"Corpus"	Chars	Diacs	Uniq. Diacs	RS	DTS	DSS
El niño bebió café en la mañana	25	4	2	0.28	0.15	0.11
הַיֶּלֶד שָׁתָה קָפֶה בַּבֹּקֶר	14	14	6	0.33	0.54	0.52

Table 1: Example sentences in Spanish and Hebrew illustrating how diacritic-based metrics are computed. The table reports the number of characters, total diacritics, number of unique diacritics, and the resulting RS, DTS, and DSS values.

metric evaluates the contribution of each individual mark.

Let d be a diacritic mark appearing as part of a rune r with base character c, We define:

$$P(d \mid c) = \frac{\#(d, c)}{\sum_{d' \in D} \#(d', c)},$$

where $\#(d, c)$ is the frequency of mark d with character c.

The diacritic token surprisal of rune r is:

$$DTS(r) = -\sum_{d \in r} \log P(d \mid c).$$

This metric captures token-level diacritic unpredictability. It is especially informative in systems where multiple diacritics can occur on the same character at the same time.

Diacritic Structural Surprisal. Diacritic structural surprisal (DSS) measures structural competition independently of frequency. Instead of asking how often a mark appears, it asks how many distinct diacritized forms it appears in.

For a base character c, let $T(c)$ denote the set of distinct runes formed from c, and let $T_d(c) \subseteq T(c)$ denote the subset of those runes that contain diacritic d. We define:

$$P_\delta(d \mid c) = \frac{|T_d(c)|}{|T(c)|}.$$

The structural surprise of rune r is:

$$DSS(r) = -\sum_{d \in r} \log P_\delta(d \mid c).$$

This metric captures type-level structural complexity. In single-diacritic systems, structural and token-level effects tend to align. In multi-diacritic systems, they can diverge substantially.

Diacritic Density. As a baseline surface measure, we define *diacritic density* as the proportion of diacritic tokens relative to base character tokens in the corpus:

$$\text{Density} = \frac{\sum_d \#(d)}{\sum_c \#(c)}.$$

This metric captures the overall diacritic load of the writing system, without conditioning on character identity or structural competition. While density does not measure ambiguity directly, it serves as a baseline indicator of how heavily diacritics are used in the writing system.

4. Data

We perform experiments and statistical analysis on fifteen languages across various scripts and typological characteristics. We include Latin-derived writing systems along with scripts like Arabic, Hebrew, and Bengali. For various languages, we provide multiple corpora. Our dataset includes 24 corpora, sourced from a mix of Universal Dependencies (UD; de Marneffe et al., 2021) and extensive web corpora. All corpora are fully diacritized. To the best of our knowledge, all diacritics have been manually added to the texts by either original authors or professionals, with the exception of some portions of the Hebrew data which were semi-automatically diacritized using Dicta's Nakdan API (Shmidman et al., 2020) followed by manual correction (see Gershuni and Pinter (2022)).

To ensure comparability across languages and corpora, we implemented a fixed-size sampling approach. For every corpus, we sample roughly 300,000 characters of diacritized text. Sampling occurred at sentence level: corpus sentences were shuffled and gradually accumulated until the target character count was achieved. For corpora that are smaller than the target (e.g., Latin), all available data was used and augmented by resampling sentences as needed.

Across various scripts, the mapping between characters and diacritics differs. In certain writing systems, a rune aligns with a single Unicode code point. For example, some Latin exist as precomposed characters: á (U+00E1, 'Latin Small Letter A with Acute'). Even letters with two diacritics may be encoded as a single code point, such as ắ (U+1EAF, 'Latin Small Letter A with Breve and Acute'). In other scripts, a rune is represented as a base character followed by one or more combining marks. For example, the Hebrew rune בָּ consists of the base character U+05D1 (HEBREW LETTER BET) together with U+05BC (HEBREW

Language Source	Diacritic Density %	Multi Diacritics %	% Words	% Lines	Mean Diacs/Word	# Runes	System
German	1.357	0.000	8.011	67.021	1.017	3	Single
UD 2.15 (GSD), 2.10 (HDR)							
Spanish	2.336	0.000	11.349	86.454	1.008	7	Single
UD2.10 (AnCora), 2.9 (GSD)							
Croatian	2.607	0.000	13.812	73.601	1.052	4	Single
Náplava et al. (2018)							
Galician	2.984	0.000	15.894	71.931	1.003	6	Single
UD 2.12 (CTG)							
Portuguese	3.700	0.000	16.229	64.003	1.140	12	Single
UD 2.13 (GSD, Cintil)							
French	3.845	0.000	17.539	90.684	1.119	13	Single
UD 2.7 (GSD), 2.2 (FTB)							
Romanian	5.999	0.000	29.488	84.026	1.141	5	Single
UD 2.8 (RRT, SiMoNERo)							
Turkish	6.510	0.000	30.971	74.863	1.347	9	Single
UD 2.8 (Penn, Kenet)							
Lithuanian	7.155	0.000	39.789	94.730	1.010	9	Single
UD 2.8 (Alksnis)							
Latin	11.118	0.000	53.338	96.053	1.207	6	Single
Chapters 1–6 of Vergil's Aeneid from Pharr's reader, digitized by Kyle Gorman							
Czech	14.505	0.000	55.119	93.514	1.510	15	Single
UD 2.15 (CAC, PDTC)							
Vietnamese	25.216	10.200	83.090	99.571	1.480	64	Multi
Náplava et al. (2018)							
Bengali	58.907	4.826	92.467	100.000	2.439	390	Multi
Leipzig (Goldhahn et al., 2012): Bengali News 2020 100k, Wiki 2021 100k							
Hebrew	66.243	14.096	99.823	99.761	3.607	325	Multi
Nakdimon (Gershuni and Pinter, 2022): Lit, news							
Arabic	82.813	9.297	99.552	89.221	3.686	353	Multi
Tashkeela (Zerrouki and Balla, 2017)							

Table 2: Corpus-level diacritic usage across languages. For languages represented by multiple corpora, statistics are averaged across corpora. The columns present overall diacritic density (ordering key), proportion of multi-diacritic characters, percentage of words and lines that are diacritized, mean diacritics per **diacritized** word, corpus-wide number of distinct character–diacritic combinations (runes), and classification as single- or multi-diacritic writing systems.

POINT DAGESH OR MAPIQ) and U+05B8 (HEBREW POINT QAMATS). Although this sequence is encoded as three unicode code points (בּ + ◌ּ + ◌ָ), it functions as a single letter with diacritics.

For scripts where diacritics are encoded as combining marks (Arabic, Hebrew, Bengali), we normalize all text to a consistent decomposed representation and compute statistics with respect to the underlying base character. Importantly, we treat each base character together with its associated diacritics as a single orthographic unit (rune), rather than as separate characters, even though they are encoded as multiple code points in Unicode.

4.1. Diacritic Usage Statistics

The statistics in Table 2 characterize the frequency, distribution, and combinatorial diversity of diacritics across writing systems. However, they do not directly capture predictability or ambiguity in character-diacritic mapping, which we aim to quantify using the information theoretic metrics introduced in Section 3.

5. Experimental Setup

We evaluate two established diacritization architectures in order to examine how properties of writing systems relate to model behavior across scripts. Our goal is not to propose a new modeling approach, but to analyze performance variation as a function of script-level characteristics.

BERT-based Diacritizer. We use the BERT-based (Devlin et al., 2019) sequence labeling model introduced by Náplava et al. (2021), which has been shown to achieve strong performance on Latin-script languages. Separate models were trained for each corpus in our dataset. For Latin-based scripts, we apply the original imple-

mentation without modification. For non-Latin scripts (Arabic, Hebrew, Bengali), we adapted the pipeline to work with writing systems in which the diacritics are encoded as separate combining marks rather than as precomposed characters. The transformer architecture, tokenization strategy (character-level), and training regime remain as in the original work.

For languages represented by multiple corpora, we additionally conducted cross-corpus evaluation. The model was trained on one corpus and evaluated on a different corpus from the same language in order to check for domain robustness.

RNN-based Diacritizer. For Latin-script languages only, we evaluate the character-level RNN model of Náplava et al. (2018), designed specifically for Latin alphabets where diacritics are precomposed in Unicode. Its encoding assumptions do not extend to Arabic, Hebrew and Bengali, so we restrict it to Latin-script corpora.

Both models use the original training objectives and decoding strategies from the reference implementations.

5.1. Evaluation Metrics

We evaluate diacritization performance using two complementary accuracy measures:

- **Word-level accuracy**, defined as exact match of full word's diacritization.
- **Rune-level accuracy**, defined as correct restoration of individual character–diacritic combinations.

Word-level accuracy reflects end-user correctness in downstream applications, while rune-level accuracy provides a more fine-grained view of model behavior and error patterns within words.

6. Results

6.1. Correlation Between Diacritic Properties

We first examine the inter-correlation between the proposed theoretical metrics. For each corpus, we compute the mean token-level complexity score for each metric and present it in Table 3.

Across all corpora, the metrics are extremely highly correlated ($|r| > 0.97$ for all pairwise comparisons), suggesting that at a broad cross-linguistic level they capture a shared underlying dimension of orthographic complexity. However, when separating the languages by script type, clearer structural differences emerge.

In multi-diacritic scripts, Rune Surprisal (RS) and Diacritic Token Surprisal (DTS) remain strongly correlated ($r = 0.93$), since both are based directly on observed corpus frequencies. Diacritic Structural Surprisal (DSS) is almost perfectly correlated with diacritic density ($r = 0.987$). This means that in these systems, languages with more diacritics overall also tend to allow more distinct diacritic combinations. Frequency-based measures correlate only moderately with structural measures ($r \sim 0.75$), suggesting that unpredictability of individual diacritics and the diversity of combination patterns are related but not identical properties.

In single-diacritic scripts, all metrics remain uniformly high ($r > 0.83$), with RS and DTS nearly identical ($r = 0.98$). This reflects the fact that when only one diacritic may attach to a character, frequency-based and structural measures collapse into a single dominant dimension.

6.2. Model Performance as a Function of Diacritic Properties

We examine the relationship between corpus-level diacritic metrics and diacritization accuracy across all languages and corpora using the BERT-based model. Table 4 reports Pearson correlations between each complexity metric and both word-level and rune-level accuracy.

Across all languages, word-level and rune-level accuracy for the BERT-based diacritizer shows a strong and highly significant negative correlation with diacritic complexity, which indicates that languages with higher orthographic complexity tend to score lower on restoration accuracy. This pattern holds across all metrics. Restricting the analysis to Indo-European languages show the same overall trend. This suggests that the effect is not limited to cross-family differences but also remains within a single language family.

Single-Diacritic Scripts. We next restrict the analysis to single-diacritic scripts, comprising 11 languages and 17 corpora. As observed above, in these scripts, the statistical metrics are mostly aligned.

For the BERT model, at both word and rune-level, DTS and DSS metrics show moderate correlations with accuracy. At rune-level, the Rune Suprisal metric is also moderately correlated. However, the overall effect is smaller and less significant than all languages.

In contrast, the RNN model exhibits strong and consistent negative correlations between diacritics complexity and accuracy across all metrics. At the word-level, correlations remain moderate, with RS showing the strongest association. This pattern suggests that the RNN architecture is more sensitive than BERT to increases in complexity even

Language	Family	Density	RS	DTS	DSS	RNN-Word	RNN-Rune	BERT-Rune	BERT-Word
German	Indo-Euro	1.374	0.044	0.031	0.009	94.245	99.088	99.045	94.888
Spanish	Indo-Euro	2.330	0.092	0.069	0.018	91.513	98.426	98.299	94.038
Croatian	Indo-Euro	2.583	0.058	0.039	0.023	93.753	98.911	99.336	96.613
Galician	Indo-Euro	2.985	0.111	0.082	0.021	92.488	98.717	99.142	95.963
Portuguese	Indo-Euro	3.694	0.150	0.114	0.047	94.805	99.033	99.170	96.526
French	Indo-Euro	3.763	0.131	0.096	0.058	91.539	98.405	97.627	91.421
Romanian	Indo-Euro	6.043	0.172	0.118	0.056	90.492	98.291	98.957	94.913
Turkish	Turkic	6.536	0.122	0.071	0.057	93.939	98.831	99.284	96.338
Lithuanian	Indo-Euro	7.127	0.193	0.133	0.068	99.659	97.335	97.048	85.726
Latin	Indo-Euro	11.181	0.237	0.145	0.072	90.140	98.320	98.957	94.729
Czech	Indo-Euro	14.434	0.327	0.207	0.121	77.613	95.567	97.052	87.353
Vietnamese	Austroasiatic	25.234	0.783	0.586	0.513	67.787	92.369	91.475	74.152
Bengali	Indo-Euro	58.635	1.593	1.207	1.168	–	–	80.668	60.770
Hebrew	Afro-Asiatic	66.026	1.706	1.366	1.487	–	–	77.996	54.479
Arabic	Afro-Asiatic	82.017	1.430	1.350	1.784	–	–	67.064	35.637

Table 3: Language-level averages of diacritic complexity metrics and diacritization performance, ordered by increasing diacritic density. For languages represented by multiple corpora, values are averaged across corpora. RS = Rune Surprisal; DTS = Diacritic Token Surprisal; DSS = Diacritic Structural Surprisal. Model performance is reported as word-level and rune-level accuracy for RNN and BERT-based diacritizers.

when structural variation is limited to a single diacritic per character.

Multi-Diacritic Scripts. We next examine multi-diacritic scripts, comprising 4 languages and 6 corpora, numbers which affect the power of our analysis. In these systems, base characters may carry multiple simultaneous diacritics, and frequency-based and structural measures diverge.

For the BERT model, a clear pattern emerges at the word and rune level despite the small sample size. Structural measures show substantially stronger negative associations with accuracy than frequency-based measures. In contrast to the global and single-diacritic analyses, where the metrics largely behave similarly, structural complexity becomes the dominant predictor of performance. Although the analysis includes only six corpora, the direction of the effects is consistent with the findings so far.

This pattern suggests that, in multi-diacritic scripts, BERT performance is more strongly affected by structural complexity than by token-level unpredictability alone. In other words, the model is more sensitive to how many structurally distinct combinations there are than to how skewed the distribution of diacritics is.

Overall, the results indicate that multi-diacritic scripts introduce a qualitatively different form of orthographic complexity, which reveals that structural combinatorics plays a distinct and predictive role in model performance.

6.3. Cross-Corpus Evaluation

We examine cross-domain robustness within individual languages. All non-Hebrew corpora were drawn from Universal Dependencies (UD; de Marneffe et al., 2021), while the Hebrew corpora were taken from the fully diacritized dataset introduced by Gershuni and Pinter (2022).

Cross-corpus evaluation was conducted for the following language pairs: Hebrew (literary / news), French (FTB, news / GSD, web), Spanish (AnCora, news / GSD, web), Turkish (Kenet, grammar examples / Penn, news), Romanian (RRT, news and web / SiMoNERo, medical), and Portuguese (CINTIL, mixed genres / GSD, web). Across languages, cross-corpus evaluation shows a modest decrease in performance relative to in-domain results. On average, rune-level accuracy drops by approximately 1.5%, and word-level accuracy by around 3%. The drop is reasonable, as different domains can contain unique vocabularies and may vary in the distribution of character-diacritic combinations. These domain differences can affect restoration accuracy, particularly given the relatively limited size of each sampled corpus (approximately 300K characters) At the same time, the relatively small decrease suggests that diacritic usage is largely consistent within a language. Although specific words change across corpora, the patterns remain stable, allowing the model to generalize well across domains.

7. Conclusion

We presented the first comparative study of the orthographic phenomenon of diacritization, ana-

Regime	Metric	BERT-Rune	BERT-Word	RNN-Rune	RNN-Word
Single-diacritic scripts					
	RS	−0.47	**−0.55***	**−0.80*****	**−0.62****
	DSS	**−0.50***	**−0.57***	**−0.76*****	**−0.57***
Multi-diacritic scripts					
	RS	−0.59	−0.50	–	–
	DSS	**−0.95****	**−0.92****	–	–
Indo-European					
	RS	**−0.98*****	**−0.97*****	–	–
	DSS	**−0.99*****	**−0.96*****	–	–
All Languages					
	RS	**−0.94*****	**−0.94*****	–	–
	DSS	**−0.99*****	**−0.98*****	–	–

Table 4: Pearson correlations between diacritization accuracy and diacritic complexity metrics: Rune Surprisal (RS) and Diacritic Structural Surprisal (DSS). Significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

lyzed through statistical and information-theoretic constructs pertaining to the intrinsic properties of diacritized text across two dozen corpora. We show that, as expected, the statistical presence and surprisal of diacritic marks across texts in a language affects the ability of state-of-the-art off-the-shelf diacritization models to restore marks in that language. In future work, we will act on our actionable conclusions and provide practical recommendations for diacritics restoration and for handling undiacritized text in downstream NLP systems, building also on the principles demarcated by Gorman and Pinter (2025).

An additional avenue for future work would be to enrich the presented study by incorporating additional features for correlation analysis, such as the functional role of diacritics in a language. For example, diacritics may encode phonological, morphological, or tonal information; these differences can help explain variation in model performance beyond structural complexity alone.

Acknowledgements

We would like to thank Kyle Gorman and the anonymous reviewers for valuable comments on earlier drafts. This research was supported by grant no. 2022215 from the United States–Israel Binational Science Foundation (BSF), Jerusalem, Israel.

8. Bibliographical References

Yonatan Belinkov and James Glass. 2015. Arabic diacritization with recurrent neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2281–2285, Lisbon, Portugal. Association for Computational Linguistics.

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.

Kareem Darwish, Hamdy Mubarak, and Ahmed Abdelali. 2017. Arabic diacritization: Stats, rules, and hacks. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 9–17, Valencia, Spain. Association for Computational Linguistics.

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

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.

Elazar Gershuni and Yuval Pinter. 2022. Restoring Hebrew diacritics without a dictionary. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1010–1018, Seattle, United States. Association for Computational Linguistics.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection:

From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Kyle Gorman and Yuval Pinter. 2025. Don't touch my diacritics. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 285–291, Albuquerque, New Mexico. Association for Computational Linguistics.

Cvetana Krstev, Ranka Stanković, and Duško Vitas. 2018. Knowledge and rule-based diacritic restoration in Serbian. In *Proceedings of the Third International Conference on Computational Linguistics in Bulgaria (CLIB 2018)*, pages 41–51, Sofia, Bulgaria. Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences.

Jakub Náplava, Milan Straka, and Jana Straková. 2021. Diacritics Restoration using BERT with Analysis on Czech language. *The Prague Bulletin of Mathematical Linguistics*, 116:27–42.

Jakub Náplava, Milan Straka, Pavel Straňák, and Jan Hajič. 2018. Diacritics restoration using neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Avi Shmidman, Shaltiel Shmidman, Moshe Koppel, and Yoav Goldberg. 2020. Nakdan: Professional Hebrew diacritizer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 197–203, Online. Association for Computational Linguistics.

Reut Tsarfaty, Shoval Sadde, Stav Klein, and Amit Seker. 2019. What's wrong with Hebrew NLP? and how to make it right. 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): System Demonstrations*, pages 259–264, Hong Kong, China. Association for Computational Linguistics.

Taha Zerrouki and Amar Balla. 2017. Tashkeela: Novel corpus of arabic vocalized texts, data for auto-diacritization systems. *Data in Brief*, 11:147–151.