sustainability

MDPI

Article

Digital Linguistic Sustainability of Turkish Dialects in AI-Based Language Technologies: Technical Robustness, Representational Justice, and Educational Inclusion

Yelda Yeşildal Eraydın [1,*] and Cemile Uzun [2]

[1] Department of Turkish Language and Literature, Faculty of Humanities and Social Sciences, Fırat University, Elazığ 23119, Türkiye
[2] Turkish Language Teaching Application and Research Center (TÖMER), Fırat University, Elazığ 23119, Türkiye; cuzun@firat.edu.tr
* Correspondence: yyesildal@firat.edu.tr

Abstract

The growing use of artificial intelligence (AI)-based language technologies in learning, assessment, writing support, and speech-enabled educational platforms has made linguistic representation an increasingly important dimension of sustainable and inclusive education. This exploratory mixed-methods study examines how transcribed forms of Turkish dialects are processed within task-appropriate natural language processing (NLP) environments and how their digital representation is interpreted by linguists and dialect speakers. The study combines a spoken-language corpus compiled from 100 speakers in 14 provinces across Türkiye's seven geographical regions with spaCy- and Stanza-based morphosyntactic outputs, an exploratory BERTurk-based semantic-similarity component, and semi-structured interviews. The corpus contains approximately 60 h of recordings, 120 pages of transcripts, and 3500 annotated structures. Descriptive task-specific observations indicate recurrent processing difficulties involving compound tense forms, non-canonical word order, dialect lexicon, discourse particles, and pragmatically marked expressions. For the POS configurations examined, descriptive POS agreement was higher on standard written Turkish (SWT) reference material (89–91%) than on transcribed dialect data (60–65%). The qualitative material indicates that non-recognition may be interpreted as cultural invisibility, reduced trust in digital tools, and possible pedagogical misalignment. These educational and social dimensions are presented as potential implications rather than directly tested outcomes. The study conceptualizes digital linguistic sustainability through three interrelated dimensions: technical robustness in processing Turkish dialects, the continuity of culturally embedded linguistic knowledge, and educational inclusion.

Keywords: digital linguistic sustainability; Turkish dialects; natural language processing; inclusive artificial intelligence; sustainable education; educational technology; representational justice; linguistic diversity

Check for updates

Academic Editor: David González-Gómez

Received: 17 July 2026
Revised: 30 August 2026
Accepted: 1 September 2026
Published: 7 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Digital sustainability is concerned not only with the long-term accessibility and efficiency of technological systems but also with the capacity of digital environments to sustain social, cultural, and linguistic diversity. Language technologies now mediate access to education, information, public services, and cultural participation. Consequently, a language variety that cannot be reliably recognized, parsed, searched, or generated by digital systems

faces a new form of marginalization: it may remain vibrant in everyday speech while becoming progressively less visible in the infrastructures through which contemporary knowledge circulates.

This problem is especially consequential for dialects. NLP systems are commonly trained on corpora dominated by standardized written registers, national news, institutional texts, and centrally produced educational materials. High performance on such sources may therefore coexist with substantial failure on conversational forms, regional morphosyntax, locally embedded vocabulary, and pragmatic particles. The resulting disparity is not simply a technical robustness problem. It affects which speakers can interact successfully with digital systems, which linguistic forms are preserved in searchable data, and which varieties are treated as legitimate in AI-supported educational environments.

The United Nations Sustainable Development Goal 4 calls for inclusive and equitable quality education, while Goal 10 emphasizes the reduction in inequalities [1]. These commitments increasingly extend to digital environments, where access is shaped not only by connectivity but also by whether systems understand the language practices of their users. The United Nations Global Digital Compact similarly frames the desired digital future as inclusive, open, sustainable, fair, safe, and secure [2]. From this perspective, linguistic inclusion is part of sustainable digital development rather than a peripheral concern.

The present study examines this issue through Turkish dialects. Turkish is morphologically rich and exhibits extensive regional variation in phonology, lexicon, inflection, word order, discourse particles, and pragmatic meaning. Yet most computational resources continue to privilege SWT. Earlier Turkish dialect-recognition studies demonstrated that regional speech can be classified with machine-learning methods [3], but they generally focused on restricted speaker samples or acoustic features. Such work is valuable for identifying varieties, although classification alone does not show whether a system can interpret the linguistic structure and social meaning of regional speech.

International research has shown that NLP resources and performance are distributed unevenly across languages and varieties. Studies of linguistic diversity in NLP demonstrate that technological coverage remains concentrated in a limited set of well-resourced languages, while cross-linguistic evaluations identify systematic disparities in foundational and user-facing tasks [4,5]. Data statements have therefore been proposed to make population, variety, provenance, and intended-use assumptions explicit [6]. Fairness-oriented scholarship further cautions that "bias" should not be inferred from a performance difference alone; the relevant harm, affected speakers, and sociolinguistic hierarchy must be specified [7]. Community-centred work similarly stresses that language technology should expand speakers' capabilities and respect local authority over linguistic knowledge [8]. These perspectives establish the theoretical connection between technical coverage, representational justice, and the sustainability-oriented objectives of the present study.

However, digital preservation should not be equated with automatic standardization. Converting expressions from Turkish dialects into their SWT equivalents may facilitate retrieval, but it can also remove pragmatic force, local memory, interactional meaning, and identity. Kelly-Holmes [9] warns against reducing living language to what is computationally measurable, while Erdocia et al. [10] argue that language must be approached as a social practice rather than merely as a dataset. Digital linguistic sustainability therefore requires a balance between interoperability and continuity: systems should make dialect forms computationally accessible without treating them as defective approximations of a single norm.

1.1. A Three-Dimensional Framework for Digital Linguistic Sustainability

This study uses digital linguistic sustainability as an integrative framework comprising three interdependent dimensions.

First, technical robustness refers to a system's capacity to process regional and standard forms consistently across tasks such as tokenization, part-of-speech tagging, dependency analysis, and semantic comparison. A technically sustainable system should not lose basic functionality when users depart from the dominant written register.

Second, cultural and representational continuity concerns whether digital systems preserve regionally embedded meanings, discourse functions, and linguistic identities. Representation is sustainable when speakers' forms remain visible and interpretable rather than being erased through automatic correction or decontextualized data extraction.

Third, educational inclusion concerns the effects of language technologies on learners and teachers. AI-supported tools become pedagogically unsustainable when regional speech is repeatedly marked as wrong, unintelligible, or inferior. Conversely, culturally responsive systems can support dialect awareness, metalinguistic reflection, and equitable participation in education.

These dimensions are mutually dependent. Technical performance without cultural interpretation may produce efficient but exclusionary systems; cultural documentation without usable technological infrastructure may remain inaccessible; and educational deployment without representational safeguards may reinforce linguistic insecurity.

1.2. Sustainable and Inclusive Education as a Use Context

AI-supported education increasingly relies on language-sensitive functions, including automated writing feedback, speech recognition, intelligent tutoring, conversational assistance, reading support, and assessment. These applications are not linguistically neutral. When systems are optimized primarily for SWT, learners who use regional phonological, lexical, morphosyntactic, or pragmatic forms may receive less accurate transcription, inappropriate correction, or misleading evaluation. A technical performance gap can therefore become an educational participation gap.

From a sustainable-education perspective, inclusion requires more than access to digital infrastructure. Learners must also be able to interact with educational technologies without their habitual language being treated automatically as noise, error, or deficiency. This requirement connects the present study to SDG 4, particularly inclusive and equitable quality education, and to SDG 10, because linguistic variation can become an overlooked mechanism through which digital systems reproduce social and regional inequalities. Sustainable educational technology should distinguish between pedagogically relevant errors and legitimate linguistic variation, communicate uncertainty transparently, and allow teachers and learners to contest or contextualize automated judgments.

The study does not assume that all dialect forms should replace SWT in formal instruction. Rather, it argues for bidialectal and variation-aware educational design: systems may support acquisition of the standard variety while recognizing Turkish dialect forms accurately, explaining contrasts without stigmatization, and preserving the learner's linguistic identity. This approach positions model robustness, culturally responsive feedback, and educator oversight as mutually reinforcing conditions of sustainable AI-supported education.

1.3. Research Gap and Contribution

Existing studies have generally examined either the technical recognition and normalization of dialects or the sociolinguistic consequences of standard-language dominance. Research on Turkish NLP has also concentrated mainly on standardized written resources, treebanks, morphology, and benchmark datasets [11–14]. The relationship among task-

specific processing difficulties, speakers' interpretations, and sustainable educational technology remains less fully developed. The present study brings these strands together through a linguistically grounded and sustainability-oriented perspective. It does not seek to develop or rank NLP architectures as a conventional computational benchmark; rather, it examines how transcribed Turkish dialect forms may be processed, represented, or overlooked in contemporary language technologies.

This study addresses that gap by combining a geographically distributed spoken corpus, task-specific NLP outputs, and qualitative interview material. Its contribution lies in relating recurrent processing observations to the linguistic properties of Turkish dialects and then considering their possible significance for cultural continuity, educational inclusion, and representational justice. The computational outputs are not treated as autonomous proof of educational or social effects, and the systems are not ranked as universally superior or inferior.

The sustainability contribution is therefore substantive rather than metaphorical. Regional language data constitute a cultural resource whose continued usability depends on digital infrastructures. A system that works well only for the dominant standard may remain operational while being socially unsustainable: it transfers the costs of technological exclusion to speakers whose linguistic practices are underrepresented. Conversely, a sustainable language technology should support long-term digital access, retain culturally situated meaning, and distribute the benefits of AI-supported education more equitably. This interpretation connects the study to SDG 4 (inclusive and equitable quality education), particularly Target 4.5 on disparities and vulnerable groups, and SDG 10 (reduced inequalities), particularly Target 10.2 on social inclusion. Participatory governance is treated as a cross-cutting principle that informs all three dimensions rather than as a separate dimension of the framework.

Table 1 summarizes the sustainability framework and the empirical indicators used in the study.

Table 1. Sustainability framework and empirical indicators used in the study.

Sustainability Dimension	Analytical Indicator in This Study	Risk Identified	Relevant Development Commitment
Technical robustness	Task-specific SWT-dialect differences and recurrent processing patterns	Unequal access to functional language technology	Inclusive digital development
Cultural and representational continuity	Preservation or loss of regional morphology, lexicon, pragmatics, and speaker meaning	Digital invisibility and computational standardization	SDG 10.2; cultural participation
Educational inclusion	Interview accounts concerning correction, trust, and possible educational use	Linguistic insecurity and unequal participation	SDG 4.5; equitable quality education

1.4. Research Questions

The study addresses the following questions:

1. What task-specific differences emerge when the examined NLP environments process SWT reference material and transcribed Turkish dialect forms?
2. Which morphological, syntactic, lexical, semantic, and pragmatic patterns recur in the processing observations?
3. How do linguists and dialect speakers interpret the representation of Turkish dialects in AI-based language technologies?
4. How can the task-specific linguistic observations and qualitative themes be interpreted together within a digital linguistic sustainability framework?

5. What potential design and governance principles follow from this linguistically grounded and interdisciplinary analysis?

The research questions are addressed through complementary analytical pathways. Questions 1 and 2 draw on the linguistic examination of task-specific outputs in the transcribed dialect data; Question 3 draws on semi-structured interviews and thematic interpretation; Question 4 is considered through interpretive integration; and Question 5 is addressed through the synthesis developed in the Discussion. This alignment establishes a direct connection between the research questions, the corresponding data sources, and the interpretive stages of the study.

2. Materials and Methods

This study adopts a linguistically grounded, exploratory mixed-methods design. The speech-data pathway identifies task-specific processing differences and recurrent linguistic patterns, whereas the interview pathway provides an interpretive perspective on representation, trust, educational inclusion, and technological justice. The two pathways are integrated only at the interpretation stage through the digital linguistic sustainability framework. Convergence is noted when both forms of material point to a related representational concern; divergence is retained where participant interpretations cannot be inferred from task-specific outputs alone. The findings are interpreted as task- and data-sensitive observations within a linguistically grounded exploratory mixed-methods framework. Mixed-methods integration is presented through a joint display following established guidance on integration at the interpretation and reporting stages [15].

Figure 1 presents the workflow of the exploratory mixed-methods design.

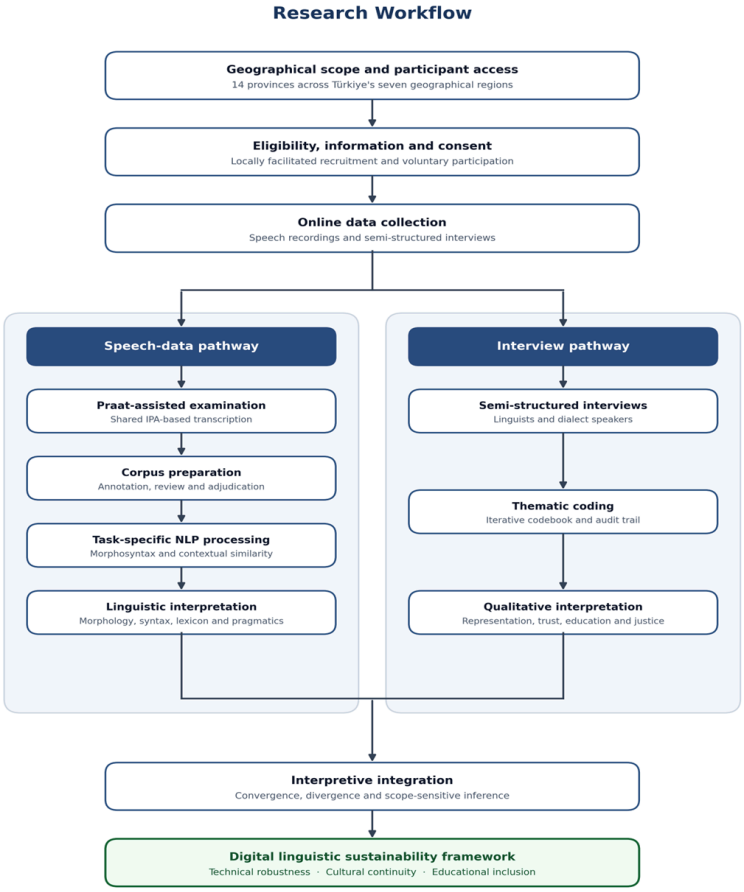


Figure 1. Research workflow for the exploratory mixed-methods design.

The workflow distinguishes the speech-data and interview pathways before their interpretive integration. The speech-data pathway proceeds from Praat-assisted examination and IPA-based transcription to corpus preparation, annotation, task-specific NLP, and linguistic interpretation. The interview pathway proceeds through semi-structured interviews and thematic coding. Their convergence and divergence are considered only at the final interpretive stage.

2.1. Data Collection and Corpus Preparation

(a) Geographical Coverage, Participant Access, and Eligibility

Oral data were collected from 14 provinces representing Türkiye's seven geographical regions: Kırklareli and Balıkesir (Marmara); Aydın and Denizli (Aegean); Adana and Mersin (Mediterranean); Konya and Kayseri (Central Anatolia); Erzurum and Ardahan (Eastern Anatolia); Şırnak and Mardin (Southeastern Anatolia); and Trabzon and Giresun (Black Sea). Two provinces were selected from each geographical region to ensure broad geographical coverage. The selection also considered the feasibility of reaching eligible speakers through academics, colleagues, and local contacts in the relevant provinces. These contacts facilitated communication with potential participants but did not take part in the recording, transcription, annotation, or analysis processes. The resulting geographical distribution provided a broad exploratory sample rather than exhaustive or statistically representative profiles of individual provincial dialects.

The speech corpus comprised 100 participants between the ages of 50 and 65, including 50 women and 50 men (mean age = 63). During recruitment, priority was initially given to speakers aged 60–65 because long-term local residence and sustained everyday language use were expected to support the retention of salient dialect features. Since this narrower age criterion could not be applied consistently across all 14 provinces, the range was extended to 50–65 years. Participants' educational backgrounds predominantly consisted of primary schooling or non-literacy, with limited representation from secondary or higher education. Eligibility required participants to have been born and raised in the relevant province and not to have permanently resided elsewhere. These criteria supported sustained exposure to the local dialect, although no individual participant was treated as representing the complete dialect profile of a province or region.

Data collection was conducted online. Prospective participants received written information explaining the purpose of the research, the voluntary nature of participation, confidentiality, the scientific use of the data, and their right to decline or withdraw. Before any questions were asked or recording began, the procedure was explained verbally, and explicit verbal informed consent was obtained. Names, signatures, and other direct identifiers were not collected. The recordings and transcripts were stored using participant codes and labelled only with age, gender, and province information.

Recordings were examined with Praat and transcribed through a shared International Phonetic Alphabet (IPA)-based protocol. The transcription retained the phonetic detail required to document linguistically relevant realizations and preserve dialect forms during their conversion into textual NLP input. The examined NLP systems received textual material derived from the transcriptions rather than acoustic recordings. The resulting analysis therefore focused on the processing of transcribed dialect forms within text-based NLP environments.

Because no existing annotated resource adequately represented the corpus, the researchers prepared and annotated the textual data. Approximately 3500 structures were classified using shared linguistic criteria informed by Universal Dependencies, including UPOS, lemma, and dependency labels where relevant. A Monte Carlo-based randomization procedure was used to organize the annotation batches and reduce possible order and

allocation effects. Two annotators independently reviewed the categorical annotations, and disagreements were resolved through adjudication. Cohen's κ values of 0.78–0.84 indicate agreement between the human annotators rather than model performance.

Evaluation examples were selected from a speaker-grouped and province-stratified subset of the annotated corpus. Material from the same speaker was kept separate from the remaining corpus material, while the provincial distribution was retained in the evaluation subset. The task-specific observations reported in this article were drawn from this subset, thereby limiting overlap across speakers and maintaining geographical coverage.

(b) Geographical Coverage, Participant Access, and Eligibility

The analytical procedure consisted of four successive stages: audio examination, transcription, human annotation, and task-specific NLP. The spaCy- and Stanza-based configurations were used to examine morphosyntactic processing through POS-related and dependency-related outputs. BERTurk-based representations were used for an exploratory examination of contextual similarity between selected dialect sentences and their SWT counterparts. The systems retained their general-purpose Turkish configurations, allowing the study to examine how existing NLP environments respond to dialect forms without prior dialect-specific adaptation. Outputs were evaluated separately according to the linguistic task supported by each system and subsequently interpreted from a linguistic perspective.

The SWT reference material was prepared according to established orthographic and morphosyntactic conventions and provided a consistent interpretive baseline for examining the selected dialect forms. Its purpose was to clarify morphological, syntactic, lexical, and contextual differences between the dialect expressions and their SWT counterparts. Because the SWT reference material and the transcribed dialect data differ in modality and register, the observed contrasts were interpreted as task- and data-sensitive patterns. Possible effects of punctuation, disfluency, speaker age, and genre were also considered. A modality-matched spoken-SWT corpus would enable a more controlled computational comparison in future research.

Contextual similarity between selected dialect–SWT sentence pairs was explored through cosine-similarity scores obtained from BERTurk-based representations. The reported values were used to identify sentence pairs requiring closer linguistic examination, particularly where dialect morphology, lexical choice, discourse particles, or pragmatic meaning affected contextual alignment. No inferential statistical significance tests were conducted on these scores; accordingly, the values are not interpreted as demonstrating statistically significant differences among sentence pairs. A fully reproducible semantic evaluation would require a separately documented protocol specifying the model checkpoint, embedding layer, pooling strategy, pair-validation procedure, calibration settings, and complete analysis code. Within the exploratory scope of the present study, the scores are used solely as descriptive indicators to guide the close linguistic examination of selected sentence pairs.

2.2. Semi-Structured Interviews

In the second phase, focused semi-structured video interviews were conducted with 28 participants from two complementary groups: 14 linguists and 14 dialect speakers. These groups were included to bring together professional assessments and speakers' experiences concerning the representation of Turkish dialects in AI-based language technologies.

The linguists were recruited through the researchers' professional and academic networks. Participation was voluntary and based on their willingness to discuss the subject. No additional quotas were applied regarding age, gender, academic seniority, institutional affiliation, or geographical location.

The dialect-speaker group comprised 14 participants, with one speaker recruited from each province included in the study. These participants were selected independently of the speech corpus, and none had contributed to the recordings used in corpus construction. A criterion-based purposive sampling strategy was used to recruit locally rooted speakers with sustained experience of the relevant dialect [16]. Priority was given to women aged 60–65 who had been born and raised in the relevant province and had not permanently resided elsewhere. Where an eligible woman could not be reached, a male participant within the same age range and with the same residence history was recruited. This selection produced a geographically distributed interview group whose members had sustained experience of local dialect use.

Each interview lasted approximately 5–8 min. The interviews followed a focused structure organized around four interrelated areas: (1) perceptions of dialect representation in educational settings; (2) experiences with and awareness of dialect processing in AI-based tools; (3) cultural belonging and perceptions of linguistic exclusion; and (4) trust in digital language technologies and patterns of use. These thematic areas provided a common framework for both participant groups, while follow-up prompts were adapted to the participants' professional knowledge or lived linguistic experiences.

All interviews were conducted in accordance with the approved ethical protocol. Participants were informed about the purpose of the research, the voluntary nature of participation, confidentiality, and the intended scientific use of the material. Recording began after explicit verbal informed consent had been obtained. The recordings were subsequently transcribed, anonymized, and prepared for thematic content analysis. All interview quotations presented in English were translated from Turkish by the authors.

2.3. Thematic Content Analysis

The interview transcripts were examined through thematic content analysis. The coding procedure combined deductive categories derived from the digital linguistic sustainability framework with inductive refinement based on the participants' accounts. Through iterative reading and comparison, the codebook was organized around four higher-order themes: (1) trust in language technologies; (2) cultural visibility and identity; (3) possible educational misalignment; and (4) technological justice and participation. Access and equality, linguistic diversity, and ethical responsibility were retained as cross-cutting codes that informed more than one thematic area.

One researcher conducted the qualitative coding manually through successive rounds of reading, initial coding, category refinement, re-coding, and analytic memoing. A stabilized codebook and an audit trail were maintained to document coding decisions and subsequent revisions. Accounts that differed from the dominant thematic tendency were retained during interpretation, enabling convergent and divergent perspectives to remain visible within the analysis.

The qualitative coding procedure was analytically separate from the linguistic reference-annotation process described in Section 2.1. The linguistic annotations provided categorical reference material for the task-specific examination of model outputs, whereas thematic coding was used to interpret participants' accounts of representation, trust, education, and technological justice. Consistency in the qualitative analysis was supported through iterative re-coding, the stabilized codebook, the audit trail, and consideration of accounts that did not fully converge with the principal interpretive patterns.

2.4. Interpretive Comparative Approach

In the final stage, the task-specific linguistic observations and qualitative themes were brought together through an interpretive joint display [15]. The comparison examined

three principal relationships: (1) recurrent morphosyntactic processing difficulties in relation to participants' accounts of trust in language technologies; (2) standard-centred outputs in relation to perceptions of linguistic visibility and cultural representation; and (3) dialect-related educational concerns in relation to prospective principles for variation-aware technological design.

The joint display was used to identify points of convergence, divergence, and complementary interpretation between the two analytical pathways. Linguistic observations indicate the forms and contexts in which processing difficulties recur, while participant accounts clarify how technological representation may be experienced and interpreted. Their integration provides a basis for discussing the potential implications of the observed patterns for digital linguistic sustainability, educational inclusion, and representational justice while preserving the distinct evidential contribution of each dataset.

3. Results

3.1. Task-Specific Processing Patterns in Transcribed Dialect Data

Research Questions 1 and 2 are addressed through task-specific linguistic observations. The morphosyntactic subsections present outputs from the spaCy- and Stanza-based configurations, whereas the BERTurk-based observations are reported in the exploratory contextual-similarity subsection. Research Question 3 is addressed through the interview themes in Section 3.1.5, and Research Question 4 through the interpretive integration presented in Section 3.1.6. Numerical agreement values, error counts, and contextual-similarity scores are reported according to their respective analytical functions.

3.1.1. Morphological Analysis (POS Tagging)

POS-related observations were obtained by comparing the spaCy- and Stanza-based outputs with the human reference annotations for the evaluation subset. The analysis focused on compound tense constructions, clipped or fused verb forms, dialect inflections, interrogative–copular combinations, and context-sensitive items whose grammatical functions depend on their use within the sentence.

The reported descriptive POS agreement values on the SWT reference material were 91% for spaCy and 89% for Stanza. The corresponding values on the transcribed dialect material were 65% and 60%, respectively. Disagreements with the human reference annotations occurred particularly in dialect verb inflections, fused grammatical structures, and discourse-sensitive items.

Illustrative POS-related tagging difficulties are presented in Table 2.

Table 2. Examples of POS-related tagging difficulties in the examined outputs.

Sentence	Model	Observed Tagging Difficulty
He gidecen mi?	spaCy	The discourse-sensitive item he was assigned an adverbial category.
Ne ediyon sen yine?	spaCy	The dialect verb form ediyon was assigned a nominal category.
Ondan mıdır nedir?	Stanza	The form mıdır was assigned an adverbial category rather than an analysis reflecting its interrogative–copular structure.

Note: Approximate English meanings: He gidecen mi? ("So, are you going?"); Ne ediyon sen yine? ("What are you doing again?"); Ondan mıdır nedir? ("Perhaps that is why.").

The examples indicate that POS assignment becomes less stable when grammatical function depends on discourse position, fused morphology, or a dialect-specific inflection. In *He gidecen mi?* the function of he is determined by its contribution to the utterance rather than by an isolated lexical category. In *Ne ediyon sen yine?* the clipped progressive and person morphology of ediyon obscures its verbal structure. The form mıdır similarly

requires sentence-level interpretation because its categorization depends on the interaction between morphology, syntax, and discourse function.

Error inspection associated the observed contrast primarily with compound tense constructions, clipped suffixes, dialect verb inflections, interrogative–copular combinations, and discourse-sensitive forms. These findings identify the principal linguistic contexts in which the examined configurations diverged from the human reference annotations.

3.1.2. Syntactic Analysis (Dependency Parsing)

Dependency-related observations were obtained from the spaCy- and Stanza-based configurations. The analysis examined predicate relations, flexible constituent order, compound or fused verb forms, and discourse-sensitive elements in the transcribed dialect material.

Illustrative dependency-related processing difficulties are presented in Table 3.

Table 3. Illustrative dependency-related difficulties in the spaCy- and Stanza-based outputs.

Sentence	Model	Observed Dependency Difficulty
Yetişemedi daha sabah.	spaCy	daha was linked to the predicate as a subject rather than being interpreted in its temporal–adverbial function.
Ben onu dün gördüm gibi.	spaCy	gibi was not linked to the predicate in a manner reflecting its sentence-level function.
Geleydin iyiydi.	Stanza	geleydin remained isolated, and no dependency relation was established with the predicate structure.

Note: Approximate English meanings: Yetişemedi daha sabah. (“It was still morning; s/he could not make it.”); Ben onu dün gördüm gibi. (“I think I saw him/her yesterday.”); Geleydin iyiydi. (“It would have been good if you had come.”).

The examples show that dependency relations are particularly sensitive to the interaction between flexible constituent order and context-dependent grammatical function. In *Yetişemedi daha sabah*, the position of *daha* contributes to an incorrect subject relation. In *Ben onu dün gördüm gibi*, the interpretation of *gibi* depends on its relation to the complete proposition. The conditional-past form *geleydin* illustrates how fused verbal morphology can interrupt the expected dependency structure.

Across the examined items, dependency-related difficulties clustered around non-canonical ordering, compound predicates, fused verb morphology, and elements expressing emphasis, stance, or sequencing. These patterns complement the POS findings by showing that dialect morphology can affect both local category assignment and sentence-level dependency relations.

3.1.3. Contextual Similarity and Linguistic Interpretation

BERTurk-based representations were used to explore contextual similarity between selected dialect sentences and their SWT counterparts. The analysis focused on the relative alignment of sentence pairs and on the linguistic features associated with higher or lower cosine-similarity scores.

Structurally close sentence pairs tended to receive higher scores within the selected examples, whereas fused morphology, dialect lexicon, idiomatic phrasing, and discourse-pragmatic functions were associated with lower alignment. The scores were therefore examined alongside the linguistic structure and contextual meaning of each sentence pair.

Table 4 presents the discourse-pragmatic functions considered during the interpretation of contextual similarity. The contribution of *gari*, *hele*, and *yoksa* extends beyond their isolated lexical meanings and emerges through their position and function within the utterance.

Table 4. Contextual-pragmatic functions of selected dialect expressions.

Dialect Sentence	Focal Expression	Contextual-Pragmatic Function
İşini hallediver de çıkalım gari.	gari	Marks urgency, conclusion, or encouragement to complete an action.
Dur hele, daha bir anlat.	hele	Organizes the sequence of actions and expresses a request for patience.
Kırdın mı kalbimi sen yoksa?	yoksa	Contributes to an indirect and affectively marked question.

The contextual-similarity scores and corresponding linguistic characteristics are reported in Table 5.

Table 5. Contextual-similarity scores and linguistic characteristics of selected dialect–SWT sentence pairs.

Dialect Sentence	SWT Counterpart	Score	Linguistic Characteristic
Ne ettin sen öyle?	Ne yaptın?	0.57	The broader pragmatic range of etmek is reduced in the SWT counterpart.
Gelecen mi?	Gelecek misin?	0.51	The pair shares a direct verbal structure despite clipped future morphology.
İyi diyim de sen bilirsin.	İtirazın yoksa olur.	0.48	The dialect sentence conveys an indirect and stance-sensitive form of consent.
Benim uşağa bak hele.	Çocuğuma baksana.	0.42	The dialect meaning of uşak and the discourse contribution of hele shape the request.
Çıha bilemedik yola.	Yola çıkamadık.	0.33	Fused dialect morphology reduces formal overlap with the SWT sentence.

The selected pairs illustrate different relationships between structural similarity and contextual meaning. Gelecen mi? retains close lexical and verbal correspondence with its SWT counterpart, whereas Çıha bilemedik yola contains fused morphology that reduces formal alignment. The examples involving hele, uşak, and diyim further show that lexical correspondence alone does not preserve stance, affect, encouragement, or local discourse convention.

3.1.4. Summary of Morphosyntactic Error Patterns

The POS and dependency observations were grouped into three recurrent linguistic categories: (1) compound or fused verb morphology that remained unresolved; (2) context-sensitive function words and discourse elements assigned an inappropriate category; and (3) dependency disruptions arising from the interaction of flexible constituent order and dialect morphology.

The counts in Table 6 refer to the incorrect outputs retained in the descriptive error analysis. Stanza produced 94 such outputs, most frequently involving unresolved verb roots or fused verbal structures. spaCy produced 87, with recurrent ambiguity between discourse-sensitive elements and adverbial categories. These counts represent the reviewed error set, while the approximately 3500 annotated structures describe the broader scope of the linguistic resource.

Table 6. Descriptive incorrect-tag counts in the reviewed dialect material.

Model	Retained Incorrect-Tag Count	Most Frequent Observed Pattern
spaCy	87	Pronoun or discourse element assigned an adverbial category
Stanza	94	Verb root or fused verbal structure left unresolved

Table 7 summarizes the retained errors across several related levels.

Table 7. Linguistic analysis of selected morphosyntactic outputs.

Form	Target Analysis	Observed Output	Linguistic Source of Difficulty
geldiydi	Verb with compound past morphology	Unresolved root in the Stanza-based output	Fused tense morphology
ki	Connective or context-sensitive function word	Adverbial category in the spaCy-based output	Discourse position and sentential function
görüyon	Verb with progressive and person morphology	Nominal category in the spaCy-based output	Clipped progressive and person marking
mıdır	Interrogative clitic combined with copular material	Adverbial category in the Stanza-based output	Fusion of interrogative and copular elements

The forms *geldiydi* and *görüyon* involve tense, aspect, and person morphology; ki depends on discourse position and sentential function; and *mıdır* combines interrogative and copular material within a fused structure. Together, these patterns locate the principal morphosyntactic contexts associated with disagreement between the examined outputs and the human reference annotations.

3.1.5. Themes from Participant Interviews

Research Question 3 was examined through four interpretive themes derived from interviews with 14 linguists and 14 dialect speakers: trust in language technologies, cultural visibility and identity, possible educational misalignment, and technological justice and participation.

Trust in Language Technologies

Participants associated the non-recognition of dialect forms with reduced confidence in AI-supported language technologies. Dialect speakers particularly emphasized the experience of being misunderstood or incorrectly represented:

> "*It doesn't recognize the words we use. It makes us feel like we're speaking incorrectly.*"
>
> *(Dialect speaker, Eastern Anatolia)*

The account connects technological reliability with the system's capacity to recognize familiar local forms. Similar observations appeared in discussions of willingness to use digital language tools and confidence in their outputs.

Cultural Visibility and Identity

Participants also interpreted dialect recognition as an issue of cultural visibility. Dialect speakers described the absence of their linguistic forms from digital systems as a reduced acknowledgement of their language experience:

> "*It's as if the Turkish we speak doesn't even count as real Turkish. . .*"
>
> *(Dialect speaker, Konya)*

These accounts relate technological representation to the recognition of locally embedded vocabulary, discourse practices, and linguistic identity.

Possible Educational Misalignment

Linguists drew attention to possible tension between students' spoken dialects and the standard-language expectations embedded in automated educational tools:

> "*Sometimes students hesitate to speak in their own dialect because the system always marks it as wrong.*"
>
> *(Linguist)*

This theme concerned the possibility that automated feedback may classify legitimate dialect forms as errors, particularly when a system does not distinguish between dialect variation and the instructional conventions of SWT. Participants connected this issue with learner confidence, language awareness, and the interpretation of automated feedback.

Technological Justice and Participation

The academic accounts emphasized the inclusion of local, social, and cultural variation in the design and evaluation of language technologies:

"*These systems don't just teach language; they reveal whose language is considered valid.*"

(Linguist, Eastern Anatolia)

This theme connected linguistic recognition with participation in corpus design, annotation, evaluation, and decisions concerning the educational use of language technologies.

3.1.6. Interpretive Integration of Linguistic Observations and Interview Themes

Research Question 4 was addressed by integrating the task-specific linguistic observations with the qualitative themes through a joint display. Three principal relationships emerged: (1) morphosyntactic non-recognition corresponded with participants' concerns about technological reliability; (2) reduced contextual alignment for dialect lexicon and discourse particles paralleled accounts of cultural visibility; and (3) ambiguity involving non-canonical and context-sensitive forms was associated with concerns about automated educational feedback.

The integrated relationships between the linguistic observations and qualitative themes are summarized in Table 8.

Table 8. Joint display of linguistic observations and qualitative themes.

Linguistic Observation	Qualitative Theme	Interpretive Relationship	Scope of Interpretation
Compound and fused verb morphology	Trust in language technologies	Recurrent non-recognition corresponded with concerns about reliability.	The relationship reflects convergence between task-specific outputs and participants' accounts.
Dialect lexicon and discourse particles	Cultural visibility and identity	Reduced contextual alignment paralleled accounts of diminished linguistic visibility.	Cultural significance was interpreted through the participants' accounts.
Non-canonical order and context-sensitive forms	Possible educational misalignment	Linguistic ambiguity was associated with concerns about automated feedback and assessment.	Educational relevance was identified as a prospective area for classroom-based evaluation.
Dialect-sensitive annotation and evaluation	Technological justice and participation	Both analytical pathways highlighted the value of speaker participation and variation-aware evaluation.	The relationship generated a design priority for subsequent research and technological development.

4. Discussion

The findings address Questions 4 and 5 by bringing the two analytical pathways into a common interpretive framework. The task-specific observations locate recurrent difficulties in the processing of dialect morphology, syntax, lexis, and pragmatics, while the interviews indicate how linguistic recognition relates to cultural visibility, trust, and possible educational experiences. Considered together, these findings show that digital linguistic sustainability involves more than the technical availability of a language in digital systems. It also concerns whether linguistic variation is processed with sufficient contextual sensitivity and whether speakers can participate in digital environments without their forms of expression being routinely reduced to SWT.

4.1. Technical Robustness as a Condition of Sustainability

The lower POS agreement values observed for the transcribed dialect material suggest that performance reported for SWT resources cannot be assumed to extend uniformly to dialect data. The recurrent difficulties involve compound verb forms, clipped or fused suffixes, flexible word order, dialect vocabulary, and discourse-sensitive particles. These features are particularly consequential in Turkish because grammatical relations and distinctions are frequently encoded within morphologically complex word forms [13,14].

The findings correspond with previous research documenting unequal technological support across languages and dialects [4,5]. They also support work in dialect NLP that emphasizes the importance of morphology-sensitive processing, transparent normalization procedures, and dedicated evaluation resources [17,18]. From this perspective, technical robustness refers not simply to high aggregate performance but to a system's capacity to process legitimate linguistic variation without consistently treating it as noise, irregularity, or error.

The present findings also illustrate why model evaluation requires linguistic interpretation. A tagging or parsing difference may arise from the interaction of corpus composition, orthographic representation, tokenization, model coverage, and the grammatical characteristics of the form itself. Consequently, the significance of an output depends on the linguistic function affected and the context in which the system is used. Fairness-oriented evaluation can contribute to this interpretation by identifying the speakers concerned, the relevant use conditions, and the possible consequences of recurrent processing difficulties [6–8].

4.2. Cultural Continuity and Computational Standardization

The semantic and pragmatic observations extend the discussion beyond formal accuracy. Dialect items such as *gari*, *hele*, and *uşak* carry meanings that depend on speaker stance, interpersonal relations, discourse sequence, affect, and local communicative conventions. Standard paraphrases may preserve the general propositional content of an utterance while weakening the pragmatic or cultural information conveyed by the dialect form.

This distinction is central to cultural continuity. Digital documentation can make dialects searchable, analysable, and accessible to future generations. However, documentation based primarily on standardization may detach linguistic forms from the contexts in which they acquire social meaning. The interview material supports this interpretation: participants associated recognition with the visibility and legitimacy of their ways of speaking, rather than viewing it solely as a matter of software functionality.

Digital linguistic sustainability therefore requires contextual representation as well as formal recognition. Provenance information, pragmatic function, speaker intention, and locally embedded meanings should be considered when dialect data are prepared, annotated, and evaluated. Normalization can remain useful for particular tasks, but it should coexist with representations that retain the original dialect form and its communicative function.

4.3. Educational Inclusion and Sustainable Learning Environments

The interview themes point to possible educational consequences of standard-centred language technologies. Automated correction, assessment, or tutoring systems may classify dialect features as errors when they are unable to distinguish legitimate variation from departures from a curricular standard. In such settings, learners may encounter a system limitation as an apparently authoritative judgment about their own speech.

The present study identifies this possibility as a design concern requiring direct educational investigation. Future research involving students, teachers, classrooms, and deployed applications would be needed to determine how automated language judgments affect participation, confidence, language awareness, and learning. This distinction al-

lows the linguistic findings to inform educational inquiry without presenting prospective implications as measured classroom outcomes.

A variation-aware approach offers a productive direction for such research. Rather than replacing a dialect form automatically, an educational system could recognize the form, explain its relationship to SWT, and distinguish between contextual register choice and linguistic deficiency. This approach could support bidialectal awareness by helping learners understand how dialect and standard forms function across different communicative settings.

For example, a hypothetical variation-aware tutor encountering *Gelecen mi?* could provide the following feedback: "This is a recognized Turkish dialect form. In SWT, the same question is written *Gelecek misin?* Use the SWT form in a formal written assignment; the dialect form remains legitimate in its regional spoken context." This feedback would teach register-sensitive standard usage without presenting the learner's dialect as an error.

These considerations are relevant to SDG 4 and SDG 10 because inclusive digital education partly depends on whether language-sensitive technologies can accommodate legitimate linguistic diversity [1,19]. The connection lies in the study's identification of a representational condition that may influence access and participation. Its educational consequences should be examined through future classroom-based and user-centred research.

4.4. An Integrated Model of Digital Linguistic Sustainability

The combined findings support an integrated model comprising three interdependent dimensions:

(1) Technical robustness: Language technologies should be evaluated on geographically and socially diverse data, with corpus composition, linguistic tasks, and evaluation conditions documented transparently.
(2) Cultural continuity: Data preparation and model evaluation should preserve contextual, pragmatic, and culturally embedded meanings alongside any task-specific normalization.
(3) Educational inclusion: Language technologies used in learning environments should distinguish legitimate linguistic variation from error and support equitable participation across different speaker groups.

These dimensions are mutually reinforcing. Technical recognition has limited cultural value if pragmatic meaning is lost; culturally rich resources have limited digital influence if contemporary systems cannot process them; and educational accessibility remains incomplete when only one form of Turkish is treated as linguistically legitimate.

Participation functions as a cross-cutting principle within this model. Dialect speakers, linguists, educators, and relevant user communities can contribute to corpus construction, annotation decisions, evaluation criteria, and assessments of appropriate technological use. Such participation helps connect technical development with the linguistic knowledge and communicative priorities of the communities represented.

4.5. Contribution to Sustainable Development

This study contributes to sustainability research by positioning linguistic diversity as part of the cultural and educational infrastructure of digital transformation. It connects task-specific processing observations with questions of contextual meaning, speaker interpretation, and equitable participation. In doing so, it brings a Turkish case into wider discussions of low-resource variation, inclusive AI, and culturally sustainable language technology.

The study also identifies a conceptual connection between SDG 4 and SDG 10. Language technologies may support more inclusive educational participation when they recognize legitimate variation and communicate the limits of automated analysis. Con-

versely, standard-centred systems may reproduce existing linguistic hierarchies when their outputs are treated as neutral or universally applicable.

Dialect recognition represents one enabling condition within this broader process. Sustainable development operates at institutional and societal levels and therefore depends on educational policy, technological governance, access, teacher practices, and community participation as well as model design. The present findings identify linguistic representation as one component of this wider structure and provide directions for its investigation in educational and social contexts.

4.6. Practical Implications

The findings suggest several priorities for Turkish language technologies. Model documentation should include dialect-sensitive evaluation materials and linguistically informed error analysis covering morphology, syntax, lexis, semantics, and pragmatics. Aggregate scores alone provide limited information when recurring errors affect culturally or communicatively significant forms.

Before automated feedback, assessment, or tutoring tools are introduced into educational settings, their outputs should be examined with speakers from different geographical and social backgrounds. Interfaces can communicate uncertainty, allow users and teachers to contest or correct outputs, and distinguish dialect forms from errors relative to a specific instructional target. Teacher guidance can further clarify the relationship between dialect variation, SWT, and register choice.

Community-informed resources are also important for long-term development. Dialect speakers, linguists, and educators can contribute to decisions about data selection, transcription, annotation, evaluation, and acceptable use. Openness should be balanced with informed consent, privacy, provenance, and community authority over linguistic data. The objective is not to construct a single fixed technological representation of each dialect, but to develop adaptable systems capable of recognizing variation and preserving contextual meaning.

4.7. Limitations and Directions for Further Research

The geographical coverage of 14 provinces provides a broad exploratory view of Turkish dialect diversity across the country's seven geographical regions. The participant profile, comprising speakers aged 50–65 in the speech corpus and dialect speakers aged 60–65 in the interviews, was selected to foreground sustained experience with local speech practices. Consequently, the sample does not capture the perspectives of younger speakers, urban speakers whose repertoires reflect mobility and dialect contact, or migrant multilingual speakers. Further research involving these groups and additional provinces will complement the present profile by examining generational change, urban and migration-related variation, multilingual repertoires, and a wider geographical range.

The linguistic analysis focuses on recurrent processing patterns across the collected dialect material rather than constructing separate dialect profiles for individual provinces. This analytical orientation supports the study's principal aim of identifying forms that may present difficulties for contemporary language technologies. Province-specific features, colloquial usage, multilingual contact, speaker-level variation, and genre-related differences offer productive dimensions for subsequent corpus-based research.

SWT material functions as an interpretive reference for examining how dialect forms are processed in relation to established orthographic and morphosyntactic conventions. A future study based on matched spoken SWT and dialect corpora could extend this comparison by controlling modality, register, disfluency, age, and genre more systematically.

The computational component is organized around the linguistic interpretation of task-specific outputs. The available records support the reported POS observations, dependency-related examples, error patterns, and exploratory contextual-similarity analysis. A dedicated computational benchmark could build on these findings through versioned pipelines, fixed model checkpoints, denominator-level reporting, UAS/LAS measures, confidence intervals, and openly documented analysis code. Such an extension would enable controlled replication while addressing a different methodological objective from the linguistically grounded inquiry pursued here.

The qualitative component enables a comparative interpretation of linguists' and dialect speakers' perspectives on linguistic recognition, trust, cultural visibility, and educational inclusion. Broader survey-based or longitudinal studies could extend these findings by examining how such perspectives vary across different demographic, generational, and geographical groups.

Together, these directions would extend the present linguistic framework through generational comparison, province-level analysis, controlled computational evaluation, and direct research in educational settings. Where ethical consent and data-protection requirements permit, the de-identified materials specified in the Data Availability Statement may also support such follow-up research.

The corpus developed for the present study also provides a foundation for a planned follow-up investigation with a more narrowly defined computational focus. That study will extend the existing material with a modality-matched spoken SWT reference, province-level linguistic comparisons, version-controlled processing pipelines, task-specific metrics, and more detailed regional analyses. These components require a separate research design and analytical framework beyond the linguistically grounded and interdisciplinary scope of the present article. The planned investigation will therefore build on the current findings by examining the identified processing patterns under more controlled and computationally reproducible conditions.

5. Conclusions

This study approaches the processing of Turkish dialects by AI-based language technologies through the framework of digital linguistic sustainability. The task-specific analysis reveals recurrent processing difficulties involving dialect morphology, flexible word order, lexical items, discourse particles, and pragmatically marked expressions. The interview findings complement these linguistic observations by relating digital recognition to cultural visibility, trust, and prospective educational use. Their interpretive integration brings together the technical, cultural, and educational dimensions of dialect representation.

Sustainable Turkish language technology requires more than expanding the volume of dialect data. Transparent corpus documentation, linguistically informed evaluation, preservation of contextual meaning, speaker participation, and a clear distinction between legitimate variation and processing error are equally important. In educational settings, these principles can guide the development of systems that recognize dialect forms, communicate uncertainty, and support awareness of the relationship between dialects and SWT. Research involving students, teachers, and deployed applications will be essential for evaluating these educational implications and examining their broader relevance to SDG 4 and SDG 10.

Turkish dialects constitute living repositories of linguistic knowledge, social memory, and cultural experience. By connecting technical robustness, cultural continuity, and educational inclusion, the study offers a linguistically grounded framework for incorporating this diversity into sustainable AI development. Its principal contribution lies in demonstrating that the digital future of Turkish depends not only on the technological processing of

language but also on the capacity of language technologies to recognize and preserve the variation through which speakers express identity, belonging, and local knowledge.

Author Contributions: Conceptualization, Y.Y.E.; methodology, Y.Y.E.; resources, C.U.; data curation, C.U.; writing—original draft preparation, Y.Y.E.; writing—review and editing, C.U. All authors have read and agreed to the published version of the manuscript.

Funding: The publication of this article was supported by the Fırat University Scientific Research Projects Coordination Unit (FÜBAP), grant number İSBF.26.32.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Fırat University Social Sciences and Humanities Research Ethics Committee (protocol code 33519; 11 April 2025).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The raw audio recordings and full interview and corpus transcriptions are not publicly available because voice characteristics and contextual information may enable participant identification, and the informed-consent procedure did not include public data archiving. Subject to the approved ethics procedure and participant-consent restrictions, the following de-identified materials may be made available by the corresponding author upon reasonable request: illustrative corpus items used in the analyses and their SWT counterparts; the transcription principles applied to those items; descriptions of the annotation categories and corresponding reference labels; and the task-specific model outputs reported for the illustrative items.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. United Nations. *Transforming Our World: The 2030 Agenda for Sustainable Development*; United Nations: New York, NY, USA, 2015. Available online: https://sdgs.un.org/2030agenda (accessed on 20 July 2026).
2. United Nations. *Global Digital Compact*; United Nations: New York, NY, USA, 2024. Available online: https://www.un.org/global-digital-compact/ (accessed on 20 July 2026).
3. Işık, G.; Artuner, H. Turkish dialect recognition using acoustic and phonotactic features in deep learning architectures. *Bilişim Teknol. Derg.* **2020**, *13*, 207–216. [CrossRef]
4. Joshi, P.; Santy, S.; Budhiraja, A.; Bali, K.; Choudhury, M. The state and fate of linguistic diversity and inclusion in the NLP world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 6282–6293. [CrossRef]
5. Blasi, D.; Anastasopoulos, A.; Neubig, G. Systematic inequalities in language technology performance across the world's languages. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022; pp. 5486–5505. [CrossRef]
6. Bender, E.M.; Friedman, B. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Trans. Assoc. Comput. Linguist.* **2018**, *6*, 587–604. [CrossRef]
7. Blodgett, S.L.; Barocas, S.; Daumé, H., III; Wallach, H. Language (technology) is power: A critical survey of "bias" in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 5454–5476. [CrossRef]
8. Bird, S. Decolonising speech and language technology. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), 8–13 December 2020; pp. 3504–3519. [CrossRef]
9. Kelly-Holmes, H. Artificial intelligence and the future of our sociolinguistic work. *J. Socioling.* **2024**, *28*, 3–10. [CrossRef]
10. Erdocia, I.; Smith, A.; Nguyen, H. Language is not a data set—Why overcoming ideologies of dataism is more important than ever in the age of AI. *J. Socioling.* **2024**, *28*, 20–25. [CrossRef]
11. Umutlu, E.E.; Cengiz, A.A.; Sever, A.K.; Erdem, N.Ş.; Aytan, B.; Tufan, B.; Topraksoy, A.; Darıcı, E.; Toraman, Ç. Evaluating the quality of benchmark datasets for low-resource languages: A case study on Turkish. In Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2), Vienna, Austria and Virtual, 31 July–1 August 2025; Volume 2025, pp. 471–487. Available online: https://aclanthology.org/2025.gem-1.41/ (accessed on 25 August 2026).
12. Türk, U.; Atmaca, F.; Özateş, Ş.B.; Öztürk, B.; Güngör, T.; Özgür, A. Improving the annotations in the Turkish Universal Dependency Treebank. In Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019), Paris, France, 29 August 2019; pp. 108–115. [CrossRef]

13. Kayabaş, A.; Schmid, H.; Topcu, A.E.; Kılıç, Ö. TRMOR: A finite-state-based morphological analyzer for Turkish. *Turk. J. Electr. Eng. Comput. Sci.* **2019**, *27*, 3837–3851. [CrossRef]
14. Kayadelen, T.; Öztürel, A.; Bohnet, B. A gold standard dependency treebank for Turkish. In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), Marseille, France, 11–16 May 2020; pp. 5156–5163.
15. Fetters, M.D.; Curry, L.A.; Creswell, J.W. Achieving integration in mixed methods designs—Principles and practices. *Health Serv. Res.* **2013**, *48*, 2134–2156. [CrossRef] [PubMed]
16. Palinkas, L.A.; Horwitz, S.M.; Green, C.A.; Wisdom, J.P.; Duan, N.; Hoagwood, K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Adm. Policy Ment. Health Ment. Health Serv. Res.* **2015**, *42*, 533–544. [CrossRef] [PubMed]
17. Dimakis, A.; Pavlopoulos, J.; Anastasopoulos, A. Dialect normalization using large language models and morphological rules. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, 27 July–1 August 2025; pp. 23696–23714. Available online: https://aclanthology.org/2025.findings-acl.1215/ (accessed on 20 July 2026).
18. Zampieri, M.; Malmasi, S.; Nakov, P.; Ali, A.; Shon, S.; Glass, J.; Scherrer, Y.; Samardžić, T.; Ljubešić, N.; Tiedemann, J.; et al. Language identification and morphosyntactic tagging: The second VarDial evaluation campaign. In Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018), Santa Fe, NM, USA, 20 August 2018; pp. 1–17. Available online: http://hdl.handle.net/10138/249333 (accessed on 20 July 2026).
19. Miao, F.; Holmes, W. *Guidance for Generative AI in Education and Research*; UNESCO: Paris, France, 2023. Available online: https://unesdoc.unesco.org/ark:/48223/pf0000386693 (accessed on 24 August 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.