Dialectal Filtering: Synthesizing Kurdish Corpora for Low-Resource Varieties by Utilizing "Noise" in Large Textual Data

Christian Schuler[⊤⊥], Raman Ahmad[⊢], Ānrán Wáng[⊤], Daniil Gurgurov[⊤⊥], Timo Baumann[⊣], Simon Ostermann[⊥⊤], Josef van Genabith[⊥⊤]
[⊤]Saarland University, [⊥]German Research Center for Artificial Intelligence (DFKI),
[⊢]HAW Hamburg, [⊣]OTH Regensburg
{christianschuler8989, raman.ahmad2022}@gmail.com, anran.wang@cs.uni-saarland.de,
timo.baumann@oth-regensburg.de, {daniil.gurgurov, josef.van_genabith, simon.ostermann}@dfki.de

Abstract

This work introduces a dialect-aware text filtering framework to pre-process, clean, and enhance large text corpora, creating variety-specific sub-corpora for neglected language varieties. We apply our framework to Kurdish, a language with rich dialectal diversity, which presents significant challenges for Natural Language Processing due to its low-resource status and the noisy nature of available text corpora. Leveraging lexicographic features, we assign multi-language-labels to text instances and synthesize over 130 dialect specific corpora from large "noisy" data sets containing unlabeled mixtures of Kurdish varieties, representing to our knowledge the largest collection of dialect-specific Kurdish NLP resources to date. This work contributes to the creation of low-resource language technology foundations, especially dialect-specific NLP applications. Specifically, we advance research on Kurdish languages by providing insights into the linguistic relationships among Kurdish varieties.

Keywords: low-resource NLP, Kurdish dialect continuum, non-English data curation, corpus synthesis, language varieties, dialect filtering, text corpus creation, multi-labels

1. Introduction

Kurdish is a language family of rich dialectal diversity (Gündoğdu et al., 2019) and severely limited natural language processing (NLP) resources (Maulud et al., 2023). Table 1 captures only the varieties with established ISO codes and some digital presence, yet even for these, basic NLP tools are largely absent. The vast majority of Kurdish varieties do not appear in this table at all and can be considered effectively zero-resource languages. Compounding this, the text data that does exist is largely *conflated*: large corpora drawn from the web blend dialect varieties without labeling them as such, rendering individual varieties invisible to standard processing pipelines (Bender, 2011; Joshi et al., 2020). We refer to such corpora as "noisy", due to the lack of finer language labels.

The macro language Kurdish (kur)[1] serves as an umbrella term, sometimes covering without distinction the many language varieties underneath it. Without the ability to detect and identify dialect-specific data, such varieties remain nothing more than non-identifiable noise in data sets. This is further complicated by inconsistent naming conventions across the literature, media, and language databases, where the same variety may appear under multiple identifiers, or multiple varieties may share a single label.

Despite exciting progress in Kurdish NLP (Morad et al., 2024; Ahmadi et al., 2024; Maulud et al., 2023), existing corpora still suffer from two critical problems. First, mixed varieties, especially in web-scraped data, often blend dialects without clear boundaries (Ahmadi, 2019). Second, weak language identification (LID) tools struggle to distinguish Kurdish from similar or dissimilar languages, let alone differentiate closely related varieties (Ball and Garrette, 2018). These issues hinder NLP progress for Kurdish varieties, especially for underrepresented sub-dialects. Recent advances in large-scale data curation (Penedo et al., 2024b; Li et al., 2024) offer promising filtering pipelines, but they rely on established language codes and thus cannot reach varieties at or near the zero-resource barrier.

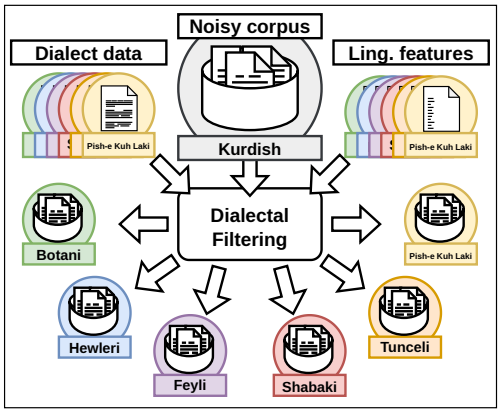

Figure 1: Dialectal filtering to derive dialect-specific text corpora from mixed data sources.

[1] https://iso639-3.sil.org/code/kur

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1505–1519
11–16 May 2026.

	Resources									Tools				
	Grammar	Corpus	UniMorph	UD	WordNet	NLLB	Wiktionary	Wikipedia	FineWeb2	LID	PoS tagger	Spell checker	ASR	MT
Northern Kurdish	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓
Central Kurdish	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Southern Kurdish	✓	✓	✓	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗
Zazaki	✓	✓	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗	✗	✗
Hawrami	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗
Laki	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗

Table 1: NLP landscape for some selected Kurdish languages, modified from (Ahmadi et al., 2025).

Our approach addresses this limitation. Rather than relying on LID tools, we leverage small but linguistically precise dialect-specific word lists as indicators to identify *dialectal noise* in large corpora and recover variety-specific text from it (see Figure 1). Our contributions are:

1. A linguistically-grounded, rule-based filtering framework for synthesizing dialect-specific sub-corpora from large, unlabeled text collections, applicable to any language with similar characteristics.

2. The largest collection of dialect-specific Kurdish NLP resources to date: 138 new variety-specific sub-corpora derived from previously conflated data, with all code publicly available[2].

2. Related Work

2.1. Kurdish Dialectology

The distinction between *language* and *dialect* is often arbitrary, influenced by sociopolitical factors, but ultimately it should suit the scope of the investigation at hand (Scherrer, 2012). *Dialect* generally refers to a variety of a language used by a particular group of people who are regarded as a single social or linguistic entity (Chambers and Trudgill, 1998). Chambers and Trudgill (1998) speak of a dialect continuum once a set of dialects are etymologically related. Languages and their speakers interact and intermingle, such that prolonged exposure to another language often leads to vocabulary borrowing and mutual influence among its native speakers (Tavadze, 2019). Some languages are known to make heavy use of loanwords from other languages (Matras, 2019) (e.g., some Kurdish dialects incorporate Arabic, Farsi and Turkish words).

Kurdish is not a monolithic language, but a collection of varieties forming a dialect continuum (Chambers and Trudgill, 1998), where adjacent dialects tend to be mutually intelligible, yet distant ones may not be (Khalid, 2020). Ozek et al. (2021) report a rather low mutual intelligibility of a Northern Kurdish and a Zazaki dialect spoken in the province of Elazığ in Turkey. This continuum, coupled with the use of multiple scripts (Latin, Arabic, Cyrillic) and influences from neighboring languages, complicates language identification and processing. Table 2 illustrates the diversity of Kurdish dialects with examples for Central Kurdish found in (Alam et al., 2024a), and collected phrases of Northern Kurdish and Hawrami. This complexity is further reflected in the inconsistent naming conventions found across the literature, media, and language databases. A single variety may carry multiple spellings of its own name and be additionally identified by the locations where it is spoken: the variety known as Shabak, for instance, appears under at least six different spellings as well as regional identifiers. Furthermore, several major dialect groups are commonly known by the name of their most prominent sub-variety (Northern Kurdish by Kurmanji, Central Kurdish by Sorani, Southern Kurdish by Palewani, Gorani by Hawrami), which effectively conceals the less prevalent varieties sheltering beneath each umbrella term. For more details about the Kurdish dialect landscape, we refer to Matras (2019)'s work that presents a rich overview of Kurdish dialect geography and a hypothetical historical scenario for how this dialect differentiation came to be.

Central Kurdish (ckb)	
Standard	من شوو ناکەم
Ardalanî	من ژن ناخوازم
Babanî	من ژن ناهێنم
Hewlêrî	ئەمن ژنی ناینم
Mukrî	ئەمن ژن ناهێنم
English	I don't marry

Northern Kurdish (kmr)	
Standard	Ez westiyayî me
Bahdînî	Ez yê westiyayî me
Efrînî	E ri'eti me
Kobanî	Ez-î westiyayî me
Qamişlokî	Ez te'bî me
Khorasani	Ez westî me
Khorasani	Ez westiya me
Central Kurdish (ckb)	
Sulaymaniyah	ماندووم
Hawrami (hac)	
Hawraman	Manîyane
English	I am tired

Table 2: Some dialectal variations in Kurdish.

[2] https://github.com/christianschuler8989/KurdishDialectFilters

2.2. Data Curation

According to Ali et al. (2025), research efforts on developing data curation pipelines for large-scale web data have soared recently (Su et al., 2025; Penedo et al., 2024a; Li et al., 2024). However, those efforts tend to rely on established language codes and rarely cover dialects, which would benefit greatly from such data curation initiatives.

Language Identification Kurdish corpora pose particular challenges for automated processing. A lack of orthographic standardization produces inconsistent spelling across sources, including variation in diacritics such as *î*, *û*, and *ê* in Latin-script Kurmanji (Esmaili et al., 2013). This compounds the difficulty for LID tools, of which recent years have nonetheless seen remarkable advances, including GlotLID (Kargaran et al., 2023) and even Kurdish-centric KurdishLID (Ahmadi et al., 2023b) and strongly related PALI (Ahmadi et al., 2023a), a LID benchmark for Perso-Arabic scripts. Kurdish-focused investigations (Hassani, 2021) and language-specific work such as for Hawrami (Khaksar and Hassani, 2024) have advanced the field. But less dominant varieties are still severely neglected.

Dialectal variations, as described above, are currently still missing from most tools and datasets. Dialect identification may be more effective when approached as a multi-label classification problem rather than the traditional single-label approach (Bernier-colborne et al., 2023). This is in line with our use of multi-labels, instead of traditional single-language labels.

Multilingual or non-English Data The extent to which text corpora can cover language varieties is directly limited by the LID tool used. To cover many languages, mmBERT (Marone et al., 2025) was additionally trained on non-English data selected from FineWeb2 (Penedo et al., 2025), which itself relies on GlotLID (Kargaran et al., 2023) for LID. The FineWeb2 dataset (Penedo et al., 2025) reports that they deliberately avoid machine learning filtering methods as these are known to disproportionately remove content in specific dialectal varieties (Blodgett et al., 2016). Similar concerns prompted us to employ a lexicon-informed rule-based approach which can later be expanded to morphological, syntactic, phonological, or other linguistic features.

2.3. Linguistic Features

The use of linguistic features has been shown to enhance NLP for under-resourced language varieties in different ways. (Nădejde et al., 2017) introduce a way to integrate syntactic information, which (Sánchez-Cartagena et al., 2020) find to be beneficial with other information such as morphological tags. (Murthy et al., 2019) report very good results for very-low-resource languages when applying syntactic reordering during transfer learning. Morphologically-aware, dictionary-based data augmentation has improved Kurdish MT (Alam et al., 2024b).

Lexicographic Features Due to the nature of the available data and with the goal of covering as many Kurdish varieties as possible, this study is limited to lexicographic features. For our use-case, lexicographic features already provide a strong foundation as related studies have shown. Our approach starts from minimal dialectal data and derives unique lexicons which aligns with efforts of weakly supervised corpus building (King and Abney, 2013; Blodgett et al., 2016). In corpus linguistics and dialectology, extracting words that are unusually characteristic of one variety compared to another is routine and some work explicitly uses these lists for classification or filtering (Rayson, 2008; Sanders, 2010). Our decision to exclude words that appear in high-resource languages (e.g. English loans) aims to avoid false positives and our unique word lists serve as a rigorous way to create lexically pure seeds that can safely be used to tag sentences as high-confidence candidates. Similar lexicon-based approaches are used in code-switching and mixed-language studies (Solorio and Liu, 2008) and lexicon-based LID (King and Abney, 2013). Lexically distinctive items (dialect-specific words, spellings, or function words) have successfully been used for dialect identification DID (Zaidan and Callison-Burch, 2014; Zampieri et al., 2017) in the past. Surveys on low-resource NLP (Pakray et al., 2025; Nekoto et al., 2020) repeatedly emphasize lexicons and dictionaries as key assets when corpora and labels are scarce. Finally, computational work on Kurdish itself often notes the scarcity and fragmentation of lexical resources, such that some projects focus explicitly on building lexicons or lexically grounded tools (Asadpour, 2023). For example, Hassani (2021) applied the Jaro measure on Swadesh lists (207 commonly used words) to calculate the similarity and difference between pairs of Kurmanji, Sorani, Hawrami, and Zazaki.

3. Methodology

3.1. Target Varieties

Initially, we collected identifiers used for Kurdish varieties in language databases (e.g., Glottolog), science, media, and via personal correspondence in an attempt to create a complete map of Kurdish

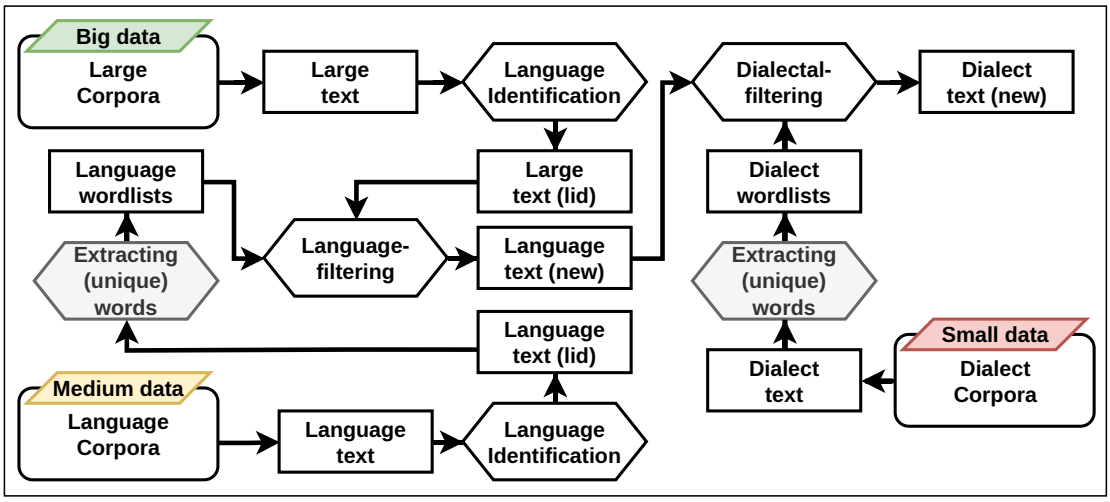

Figure 2: Simplified overview of pipeline for language and dialectal filtering of text from large corpora.

dialectology (Schuler and Ahmad, 2023). Due to different spellings and location names often used to denote varieties, such as *Mosul* for *Shabak*, we ended up with 620 different identifiers for languages that are either Kurdish varieties, or appeared in a related context. We manually aggregated these to a set of 183 varieties, for which we try to derive new text corpora, culminating in 138 new corpora.

3.2. Data Filtering

Figure 2 illustrates the three types of corpora serving as starting points from where data is moved through our pipeline. Our method of filtering variety-specific data requires linguistic features, which can be used with high confidence to identify text by being unique to a specific dialect. The current focus is on lexicographic features, which includes: variety-specific vocabulary (words unique to dialects, e.g. regional terms) drawn from **dialect corpora**.

Any word found in a new sentence, which is shared among dialects and languages, does not provide much information for identifying the exact language to which the sentence belongs to. However, using a rare word, which is known to only be found in text data from a very specific language variety, and looking it up in a text corpus, should result in matches which can be seen as strong indicators for having found text data that actually belongs to this variety.

While linguistic publications may provide linguistic features of a higher confidence and less noise, they are difficult to parse and do not yet cover many of the Kurdish varieties in such detail. Words of a language or dialect can directly be drawn from text data, if some is available. We created variety-specific word lists, for as many Kurdish dialects as possible and refined them to include only "unique" words absent from the lists of other varieties. Restraining the filtering to only consider those words that are unique in our available data seems to run counter to the spirit of reflecting a dialect continuum, as it neglects vocabulary shared across close varieties. However, this approach better reflects the high-precision character of this work and focuses on making more reliable predictions during dialect-specific data sample identification. As the resulting sub-corpora will have a higher validity, they can then inform rule refinement in an iterative development process. Later, the uniqueness constraint can be relaxed, once the aggregated dialect filter rules have been validated via e.g. human-based evaluation studies.

On the other side of this filtering method are **large corpora**, often created via web-crawling and known to contain considerable amounts of *'dialectal noise'*, i.e. a mix of unlabeled dialect varieties. This is the kind of text data which is close enough to the standard variant of a language, such that it sometimes gets identified by LID tools, or simply shares the same websites it appears on. For each line of text from a corpus, we try to find matches in our dialectal word lists. Once found, we assign language-labels, which based on the presence of unique words from multiple variety-specific lists can be more than one. This means that a sentence with unique words from multiple varieties receives multiple language variety labels, which we call multi-labels and reflect the Kurdish dialect continuum. Finally, we extract all sentences with at least one unique word from a variety into its sub-corpus (e.g., the Kobani sub-corpus includes sentences with Kobani-specific words, but the same sentences can also appear in other sub-corpora).

3.3. Multi-label Classification

Natural language use rarely aligns with clean, single-variety boundaries: a speaker may draw on vocabulary from multiple related varieties within a single sentence, particularly across a dialect continuum such as Kurdish. Rather than forcing a single language label onto each data sample, we therefore introduce a more expressive system in which multiple language varieties may be associated with one sentence, which we call *multi-labels*. Additionally, we keep track of the exact source and reasoning behind each label. This starts with the language label, which is usually provided by the original authors and publishers of the text data. These, as well as most labels, we only treat as indicators, or candidate labels, since language varieties often get mixed up due to various reasons. For new label candidates, we store additional information, such as the confidence score from LID tools (e.g. GlotLID or KLID) or, as planned in future work, the language proficiency of human annotators.

In the case of our lexicographic filtering, the exact words that were used during the process to derive the new language label candidates are tracked. This approach combines two important benefits: First, it allows people to make educated decisions about how to use the resulting text data. In cases that require high precision, e.g. due to sensitive contexts, a selection by annotator agreement and confidence extracts the most valid subset. Yet other use cases might be able to benefit even from noisy data, in which case each entry with even a single label candidate can be extracted for experiments. Second, keeping track of these details individually allows us to evaluate and re-evaluate the filtering processes down to single filter rules and data samples. For instance, knowing whether a label was assigned by an automated LID tool or by a native speaker with linguistic expertise can significantly shift a researcher's trust in that label.

4. Experiments and Results

4.1. Data Sources

Dialect Corpora Our work builds on linguistic features drawn from various sources, as linguistic databases rarely include dialectal variations. Hence, we turned to related literature e.g. Haig and Öpengin (2018)'s description on dialectal variations of Northern Kurdish, various word lists from Wikipedia pages e.g. Zimanê kurdî[3], and Matras (2016)'s Database of Kurdish Dialects.

[3] https://ku.wikipedia.org/wiki/Zimanê_kurdî

While Matras (2016) provides more than 42,000 entries from 109 different locations, it should be noted that these stem from recorded speech samples that were later transcribed, introducing noise and errors in various ways, necessitating extensive pre-processing. In addition, we draw data from relevant work (Ahmadi et al., 2025; Haig et al., 2024; Ahmadi et al., 2024; Ahmad and Schuler, 2023): PARME (Ahmadi et al., 2025) with 22, and WOWA (Haig et al., 2024), with 8 varities of Kurdish, (Ahmadi et al., 2024) with 6 varieties of Central Kurdish, and one more dataset which includes a Northern Kurdish dialect.

Language	Total Lines	Unique Words
ckb	75.0K (196.32)	87.1K (21.48)
ckb-Arab	90.9K (14.64)	77.2K (3.56)
ckb-Latn	91.7K (16.72)	71.0K (2.78)
ckb-rud	59.7K (177.04)	164.2K (5.59)
ckb-sah	77.7K (193.27)	126.3K (5.24)
ckb_Arab	3.1K (169.40)	23.8K (6.40)
ckb_Latn	815.0 (118.47)	4.2K (9.75)
hac	26.9K (601.94)	259.0K (16.77)
hac_Arab	2.4K (122.57)	17.7K (5.44)
hac_Latn	205.0 (5.02)	218.0 (4.39)
kmr	310.7K (322.23)	368.8K (30.68)
kmr_Arab	4.0K (170.61)	19.5K (9.30)
kmr_Latn	2.5K (144.33)	9.4K (9.55)
sdh	11.9K (28.33)	12.5K (9.53)
sdh-fey	80.9K (171.73)	229.8K (5.15)
sdh_Arab	2.4K (154.84)	19.4K (5.22)
zza	79.5K (123.37)	144.3K (8.90)
zza_Latn	5.0K (95.99)	22.8K (10.08)

Table 3: Data from Kurdish language corpora, average length in brackets.

Language Corpora There exist a few text corpora specifically created for Kurdish languages. For Central Kurdish this includes AsoSoft (Veisi et al., 2020), Sorani UniMorph (Ahmadi and Mahmudi, 2023), Kurdish News Dataset Headlines (KNDH) (Badawi et al., 2023), and the Kurdish Textbooks Corpus (KTC) (Abdulrahman et al., 2019). For Northern Kurdish we include the Corpus of Contemporary Written Kurdish (CCWK) (Incekan and Haig, 2021), the Corpus of Contemporary Kurdish Newspaper Texts (CCKNT) (Haig, 2001), Kurmanji UniMorph (Cotterell et al., 2017), a Dataset for Kurmanji Lemmatization and Spell-Checker (KLSC) (Hoshyar Mustafa and Nabi, 2022), and Twitter data (Kurt, 2020). Multiple Kurdish languages can be found in the Pewan corpus (Esmaili et al., 2013), (Ahmadi, 2020), (Ahmadi et al., 2019), and (Ahmadi et al., 2022) (see Table 3 for the amounts of data per language).

Large Corpora Since most large multilingual corpora omit Kurdish varieties entirely, we cast a

Language	Total Lines	Unique Words
ku	642.0K (383.99)	919.1K (23.78)
kur	14.5K (67.91)	38.2K (4.53)
ckb	2.3M (735.06)	2.8M (20.90)
kmr	610.2K (294.10)	925.4K (7.98)
ku_Latn	505.2K (60.35)	400.6K (4.59)
zza	202.0 (514.50)	6.9K (6.61)

Table 4: Kurdish languages in large text corpora, average length in brackets.

wide net by including any corpus that reports coverage of at least one Kurdish language. To serve as "Large Corpora", we include CCAligned (El-Kishky et al., 2020), mC4 (Xue et al., 2021), NLLB-200 (Team et al., 2022), OpenSubtitles [4], OSCAR (Ortiz Suárez et al., 2019), Tanzil (Tiedemann, 2012), TED2020, and wikimedia (Wikimedia, 2025) (see Table 4).

4.2. Pre-Processing

Prior to filtering, all data sources underwent corpus-type specific cleaning and normalization procedures to ensure comparability and to minimize non-linguistic noise. Across all corpus types, a shared preprocessing baseline was applied, which includes lower-casing, removal of punctuation, digits, and special symbols, whitespace collapsing, stopword removal[5], and tokenization. Word lists were additionally unified by removing entries appearing in more than one variety-specific list, reducing false positives during filtering. Due to the heterogeneous nature of the materials that ranges from highly curated dialect datasets to massive, noisy web corpora, each corpus type required tailored treatment reflecting its distinct role in the pipeline.

Data from **Dialect Corpora** were cleaned with a focus on producing linguistically accurate word lists for feature extraction with the primary aim of keeping dialectal differences. We first manually aligned inconsistent language and dialect labels wherever possible, then applied a standardized pre-processing pipeline. The cleaned text was tokenized, and unique word lists were extracted for each dialect. To isolate dialect-specific features, these word lists were cross-compared across all included dialects, and any shared vocabulary was removed. The resulting sets of unique lexical items then served as the basis for the *Dialect Filtering* stage.

[4] https://www.opensubtitles.org/en/search/sublanguageid-kur

[5] Stopword lists (de, ar, id, zh, da, en, fr, hi, it, ja, ko, ku, la, pl, pt, ru, sv, th, tr, uk, vi, yo, zu) were drawn from: https://github.com/stopwords-iso/stopwords-iso

In contrast, **Large Corpora** required a more general cleaning approach aimed at recovering usable, well-structured text rather than precise lexicon extraction. We first identified and split abnormally long words and lines using language-specific thresholds and known vocabulary. For example, a token such as "Ilikepizza" would be segmented into "I like pizza" using previously observed word boundaries. Very long text lines were divided at punctuation marks or estimated sentence-length limits to improve tokenization. We then applied LID with GlotLID (Kargaran et al., 2023) and KLID (Ahmadi et al., 2023b) to remove unrelated or erroneously assigned content (e.g., Chinese text). Only texts written in Latin, Arabic, or Cyrillic scripts were retained as potentially Kurdish. These cleaned corpora formed the input for the *Language Filtering* experiments.

Language Corpora were processed using a hybrid of the two previous strategies, balancing lexical precision with large-scale text normalization. Overly long tokens and lines were segmented as described above, and LID-based filtering was applied to remove non-Kurdish material, before the shared pre-processing pipeline was applied. From these preprocessed texts, language-level word lists were extracted, then cross-compared both across Kurdish languages and against 20 additional languages sampled from OSCAR (Ortiz Suárez et al., 2019) (see Figure 3). This multilingual cross-referencing helped identify and remove noise and loanwords, producing sets of lexically unique words for each Kurdish language, later used in *Language Filtering*.

After cleaning, all corpora were harmonized in format, yielding a unified yet type-specific text base. Dialect data provided fine-grained linguistic features for filtering; Language Corpora served as intermediate, curated references; and Large Corpora offered broad and diverse text pools from which new, dialect- and language-specific subcorpora could be extracted. This multi-level cleaning pipeline ensured that subsequent filtering and evaluation steps were based on text material that was internally consistent and minimally affected by cross-lingual noise or formatting artifacts.

4.3. Framework Application

The framework yielded 138 sub-corpora (see Table 5 for a selected subset), e.g., *Bijar* (1,261 lines of text) for which we used 15 original lines from dialectal sources. The potential of part-of-speech tagging is shown by the word *Bavê*, as in the sentence *Bavê min li malê ye* (eng: My father is at home). In a different context, and used as a verb (eng: throwing), it becomes more specific to the Kobani variety of Northern Kurdish. However these new datasets are a first step to supporting

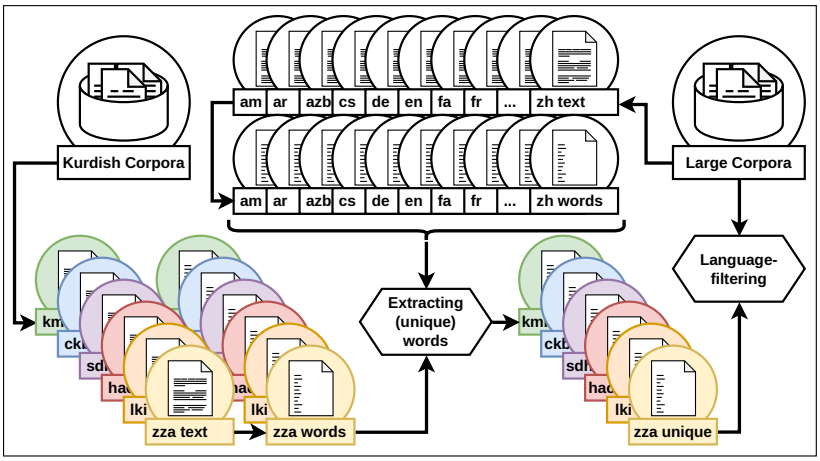

Figure 3: Cross-referencing extracted word lists for Kurdish varieties with those of many other languages allows to remove noise and loan words and to derive word lists that are unique across languages.

variety-specific NLP applications.

Variety	Original	Filtered
(sdh) Bijar	15 (32.8)	1,261 (158.8)
(zza) Kurmanjki	6 (16.2)	8820 (125.9)
(zza) Dimli	20 (43.1) +7 (17.1)	7,365 (164.4)
(sdh) Bijar (Garusi)	620 (31.9)	16,980 (256.0)
(lki) Sahneyi	213 (36.5)	6,628 (251.8)
Ad Darbasiyah	365 (16.5)	18,474 (149.2)
(hac) Gawraju	653 (34.5)	16,707 (224.0)
(kmr) Kobani	372 (16.5) +735 (48.7)	1,8 M (145.7)

Table 5: Effective growth of selected variety subcorpora for number of text lines (and avg. characters).

5. Evaluation

Evaluating the quality and linguistic characteristics of text corpora is central to low-resource NLP, and especially difficult when gold-standard benchmarks are unavailable. We assess five corpus types spanning a spectrum from raw to refined: **Large Base** (extensive, noisy), **Language Base** (curated, language-level), and **Dialect Base** (small, fine-grained) corpora serve as baselines, while **Language-Filtered** and **Dialect-Filtered** sub-corpora represent the output of our pipeline. Comparing these datasets allows us to analyze how our filtering approach shapes the linguistic character of the resulting text.

Given the absence of gold-standard reference data for Kurdish dialect identification on the level we are attempting to do, we rely on a suite of metrics that cover intrinsic corpus-level characteristics to provide an indirect but informative evaluation. We employ the Type-Token Ratio (TTR) (Templin, 1957), the Hapax Legomena Ratio (Baayen, 2001), the Zipf's Law Slope (Zipf, 1935), and the Flesch-Kincaid Grade Level (Thomas et al., 1975), computed at both word and character n-gram levels (n = 1–4). These complementary measures capture aspects of lexical diversity, rarity, distributional balance, and readability, which enables a multidimensional comparison across corpora.

Type-Token Ratio The Type-Token Ratio (TTR) measures lexical or sublexical diversity, defined as the number of unique units (types) divided by total units (tokens). For word n-grams, types are unique word sequences; for character n-grams, they are unique character sequences. A higher TTR (approaching 1) implies greater diversity, while a lower one indicates repetition or limited vocabulary.

Figure 5 and Figure 4 show TTR results at both levels. Across all n-grams, larger corpora display lower TTR values, consistent with repetitive and noisy content. The filtered corpora exhibit slightly higher TTR values than their parent corpora, positioning them between the large and dialect corpora and suggesting that our filtering process increases corpus scale while maintaining lexical diversity for our dialectal dataset.

Hapax Legomena Ratio While TTR captures diversity, the Hapax Legomena Ratio quantifies the proportion of units that appear exactly once, reflecting the prevalence of rare or unique elements. High values (near 1) indicate many unique forms, common in noisy or heterogeneous data; low values (near 0) indicate lexically constrained or formulaic text.

As shown in Figure 6, character-based Hapax Ratios behave as expected: small dialect corpora have low values for character 1-grams (reflect-

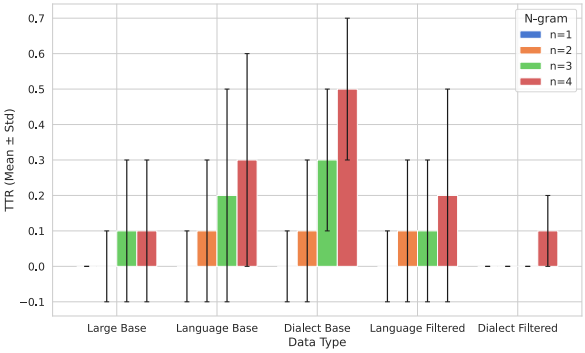

Figure 4: Character TTR by data type.

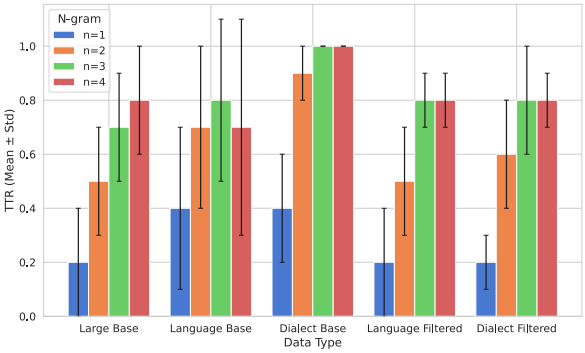

Figure 5: Word TTR by data type.

ing consistent orthography), but higher values for longer n-grams (e.g. 4-grams), revealing dialect-specific morphological sequences. Filtered corpora show ratios closer to those of large corpora, indicating that while filtering improves linguistic coherence, residual noise of larger datasets is also injected by adding new dialectal content, especially at the character level. Word-level Hapax Ratios (Figure 7) mirror these tendencies but with smaller inter-corpus variation.

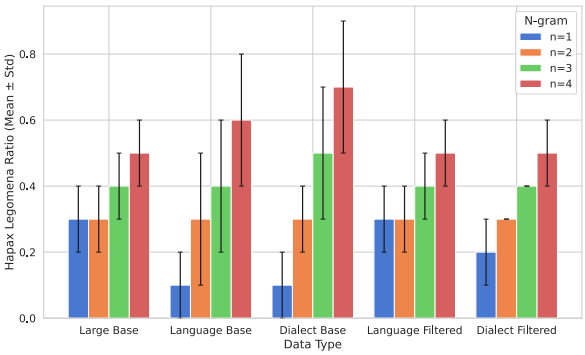

Figure 6: Character Hapax Ratio by type.

Zipf's Law Slope Zipf's Law describes the inverse relationship between frequency and rank of linguistic units. On a log-log scale, a slope near -1 indicates a natural frequency distribution typical of human language (Piantadosi, 2014). Steeper

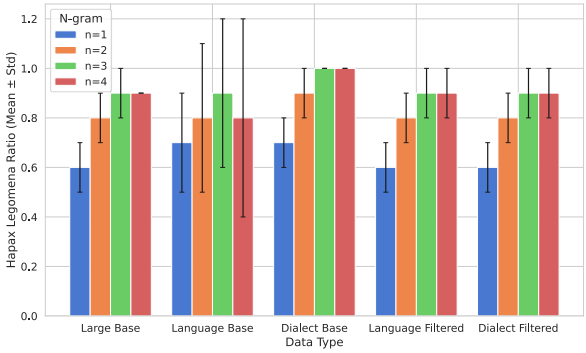

Figure 7: Word Hapax Legomena Ratio by type.

slopes (< -1) denote over-repetition or limited vocabulary, while flatter slopes (> -1) suggest uniform distributions caused by noise.

In Figure 9, large corpora cluster around the theoretical ideal of -1, while smaller dialect corpora exhibit flatter slopes, consistent with their restricted vocabularies. After filtering, both language- and dialect-filtered datasets show slopes closer to the ideal, suggesting that filtering restores linguistic balance relative to the more constrained source word lists. Character-based slopes (Figure 8) show lower values overall, reflecting rapid frequency drop-offs and residual cross-lingual noise, likely due to lenient character filtering thresholds during pre-processing.

Flesch-Kincaid Grade Level The Flesch-Kincaid Grade Level approximates text readability, representing the U.S. school grade level required to comprehend the text, based on sentence length and word complexity. Higher scores imply complex or noisy texts; lower scores reflect simpler, more colloquial language. This metric is applied only to word 1-grams.

Figure 10 aligns well with our expectations: large corpora yield the highest grade levels, reflecting complex or corrupted content; language corpora occupy intermediate positions; and dialect corpora produce the lowest values, consistent with

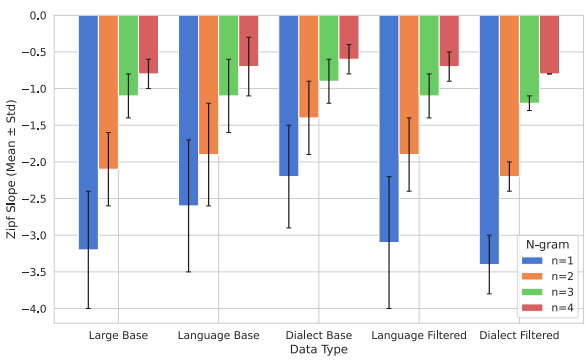

Figure 8: Character Zipf's Law Slope by data type.

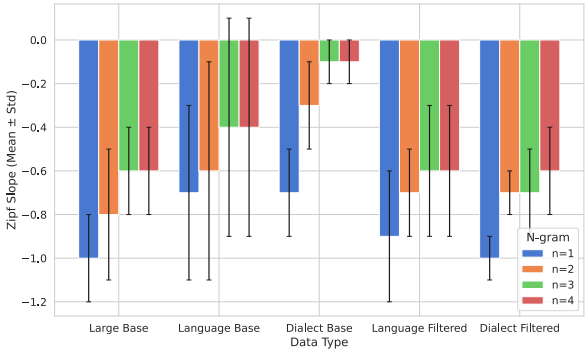

Figure 9: Word Zipf's Law Slope by data type.

their short, spoken-style text. The filtered corpora again position themselves between these extremes, exhibiting complexity levels similar to their respective data sources but slightly simplified.

Interpretation Taken together, the observed metric patterns suggest that our feature-based filtering approach can successfully extract linguistically coherent sub-corpora from large and heterogeneous text collections. Across all metrics, the *Language-Filtered* and textitDialect-Filtered corpora consistently occupy intermediate positions between the Large, Language, and Dialect Corpora. In this sense, our method not only identifies but effectively creates new language- and dialect-specific sub-corpora, enabling future research on Kurdish varieties that were previously underrepresented or completely unlabeled.

6. Discussion

The sub-corpora derived through our framework span a broad spectrum of Kurdish linguistic diversity, with many providing the first large-scale textual dataset for their respective varieties. At the same time, the evaluation results highlight that intrinsic corpus metrics, however informative, cannot fully substitute for gold-standard benchmarks or native speaker validation. The following limitations shaped several of our methodological decisions and point toward concrete directions for future work.

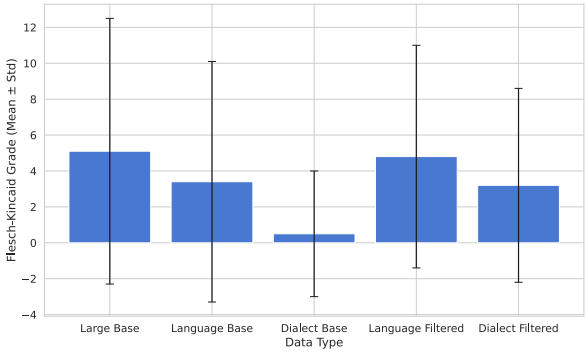

Figure 10: Flesch-Kincaid Grade Level by type.

6.1. Limitations

Kurdish Nomenclature The gaps in nomenclature for Kurdish languages present a non-trivial challenge for dialect-specific data curation. A single dialect may appear under numerous alternative labels, while conversely, multiple distinct varieties may share one identifier. Writing system ambiguity compounds this further: Latin-script text labeled as Kurdish is likely Northern Kurdish, but may equally be Zazaki or a transliterated Arabic-script variety. Notably, the Bahdînî dialect of Northern Kurdish uses Arabic script rather than Latin. As a result, some derived sub-corpora may warrant merging to better reflect how the languages actually exist, rather than how they are labeled.

Uneven and Few Data per Variety The uneven distribution of unique words across varieties in our data directly affects the number of matches accumulated during filtering. Homonyms such as *Bavê* (meaning both "father" in general Northern Kurdish and "throwing" specifically in Kobani) illustrate how ambiguous words can produce multi-label clutter and noisy assignments. Future work may address this through the integration of morphosyntactic features during filtering.

Evaluation of derived Corpora A fine-grained evaluation would ideally assess internal coherence and cross-variety distinctiveness using lexical, morphological, and syntactic features against independent benchmarks. This remains infeasible at present due to the absence of reliable dialect-level labels across available corpora. A human-based evaluation covering a meaningful range of varieties would require funding well beyond the scope of this work.

Data Curation in Low-Resourced Settings Effective curation for a dialect continuum benefits from capturing hierarchical variety relations: filtering for Kobani, for instance, would likely improve by restricting input to Northern Kurdish (kmr) data, reducing false positives from unrelated languages. However, labels in our Language Corpora often refer to broad categories (e.g., "ckb"), while Dialect Corpora use finer but inconsistent identifiers (e.g., "ckb-hwl", "Erbil", "Hewlerî"). Geographic coordinates are available for most locations, but unambiguous mapping from location to linguistic variety remains non-trivial, as many regions host speakers of multiple Kurdish varieties simultaneously.

6.2. Summary & Future Work

Despite these challenges, stable trends across 138 varieties and 12 source corpora suggest robustness: systematic patterns emerge even under individual label uncertainty, consistent with a "law of large numbers" effect. This validates our framework as a meaningful first step towards dialect-aware text extraction and motivates targeted, human-based validation studies focused on specific speaker communities.

Future efforts will standardize dialect identifiers and refine variety-to-location mappings to enable finer evaluations. We are developing an online annotation platform to involve native speakers in validating and enriching the derived sub-corpora, following participatory research principles (Bird, 2020; Nekoto et al., 2020). Methodologically, these sub-corpora will serve as seed data for machine learning classifiers targeting lexical and morphosyntactic features (Xie et al., 2024; Masis et al., 2022), alongside integration of additional linguistic dimensions and data augmentation strategies. Together, these efforts will move toward automatic dialect identification models, reducing reliance on rule-based methods and improving scalability.

7. Conclusion

Applying our framework across 12 large corpora, 14 language corpora, and five dialect corpora, we derived 138 new variety-specific sub-corpora, the largest collection of dialect-specific Kurdish text resources to date. The resulting datasets lay a foundation for dialect-aware NLP and linguistic research, supporting future benchmarking and model development (Faisal et al., 2024). Limitations due to lexical ambiguity and inconsistent labeling observed in source corpora highlight the need for morphological and syntactic feature integration and native speaker validation (Bird, 2020). Once refined, these sub-corpora will advance dialect-sensitive NLP and contribute to Kurdish language documentation and preservation.

8. Acknowledgements

The first author gratefully acknowledges financial support from GraduSaar, University of Saarland. This work was additionally funded by the German Federal Ministry of Research, Technology and Space (BMFTR) as part of the project TRAILS (01IW24005). We thank the anonymous reviewers for their thorough review and constructive feedback.

9. Bibliographical References

Roshna Omer Abdulrahman, Hossein Hassani, and Sina Ahmadi. 2019. Developing a fine-grained corpus for a less-resourced language: The case of kurdish.

Raman Ahmad and Christian Schuler. 2023. Analysis of phonology and morphology in the kobani dialect [conference poster & abstract].

Sina Ahmadi. 2019. A rule-based kurdish text transliteration system. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18(2):18:1–18:8.

Sina Ahmadi. 2020. Building a Corpus for the Zaza–Gorani Language Family. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 70–78, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Sina Ahmadi, Milind Agarwal, and Antonios Anastasopoulos. 2023a. Pali: A language identification benchmark for perso-arabic scripts. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 78–90, Dubrovnik, Croatia. Association for Computational Linguistics.

Sina Ahmadi, Zahra Azin, Sara Belelli, and Antonios Anastasopoulos. 2023b. Approaches to corpus creation for low-resource language technology: The case of southern kurdish and laki.

Sina Ahmadi, Hossein Hassani, and Daban Q. Jaff. 2022. Leveraging multilingual news websites for building a kurdish parallel corpus. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 21(5):1–11.

Sina Ahmadi, Hossein Hassani, and John P. McCrae. 2019. Towards electronic lexicography for the kurdish language. eLex 2019.

Sina Ahmadi, Daban Jaff, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2024. Language and speech technology for central kurdish varieties. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10034–10045, Torino, Italia. ELRA and ICCL.

Sina Ahmadi and Aso Mahmudi. 2023. Revisiting and amending central kurdish data on unimorph 4.0. In *Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 38–48, Toronto, Canada. Association for Computational Linguistics.

Sina Ahmadi, Rico Sennrich, Erfan Karami, Ako Marani, Parviz Fekrazad, Gholamreza Akbarzadeh Baghban, Hanah Hadi, Semko Heidari, Mahîr Dogan, Pedram Asadi, Dashne Bashir, Mohammad Amin Ghodrati, Kourosh Amini, Zeynab Ashourinezhad, Mana Baladi, Farshid Ezzati, Alireza Ghasemifar, Daryoush Hosseinpour, Behrooz Abbaszadeh, Amin Hassanpour, Bahaddin Jalal Hamaamin, Saya Kamal Hama, Ardeshir Mousavi, Sarko Nazir Hussein, Isar Nejadgholi, Mehmet Ölmez, Horam Osmanpour, Rashid Roshan Ramezani, Aryan Sediq Aziz, Ali Salehi Sheikhalikelayeh, Mohammadreza Yadegari, Kewyar Yadegari, and Sedighe Zamani Roodsari. 2025. Parme: Parallel corpora for low-resourced middle eastern languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, Vienna, Austria. Association for Computational Linguistics.

Md Mahfuz Ibn Alam, Sina Ahmadi, and Antonios Anastasopoulos. 2024a. Codet: A benchmark for contrastive dialectal evaluation of machine translation.

Md Mahfuz Ibn Alam, Sina Ahmadi, and Antonios Anastasopoulos. 2024b. A morphologically-aware dictionary-based data augmentation technique for machine translation of under-represented languages.

Mehdi Ali, Manuel Brack, Max Lübbering, Elias Wendt, Abbas Goher Khan, Richard Rutmann, Alex Jude, Maurice Kraus, Alexander Arno Weber, David Kaczér, Florian Mai, Lucie Flek, Rafet Sifa, Nicolas Flores-Herr, Joachim Köhler, Patrick Schramowski, Michael Fromm, and Kristian Kersting. 2025. Judging quality across languages: A multilingual approach to pretraining data filtering with language models.

Hiwa Asadpour. 2023. A corpus analysis of the effects of definiteness and animacy on word order variation. *Languages*, 8(4):279.

Harald Baayen. 2001. *Word Frequency Distributions*. Kluwer Academic Publishers.

Soran Badawi, Ari M. Saeed, Sara A. Ahmed, Peshraw Ahmed Abdalla, and Diyari A. Hassan. 2023. Kurdish news dataset headlines (kndh) through multiclass classification. *Data in Brief*, 48:109120.

Kelsey Ball and Dan Garrette. 2018. Part-of-Speech Tagging for Code-Switched, Transliterated Texts without Explicit Language Identification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3084–3089, Brussels, Belgium. Association for Computational Linguistics.

Emily M. Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

Gabriel Bernier-colborne, Cyril Goutte, and Serge Leger. 2023. Dialect and variant identification as a multi-label classification task: A proposal based on near-duplicate analysis. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 142–151, Dubrovnik, Croatia. Association for Computational Linguistics.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

J. K. Chambers and Peter Trudgill. 1998. *Dialectology*, 2 edition. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. Conll-sigmorphon 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzman, and Philipp Koehn. 2020. Ccaligned: A massive collection of cross-lingual web-document pairs.

Kyumars Sheykh Esmaili, Donya Eliassi, Shahin Salavati, Purya Aliabadi, Asrin Mohammadi, Somayeh Yosefi, and Shownem Hakimi. 2013. Building a test collection for sorani kurdish. In *2013 ACS International Conference on Computer Systems and Applications (AICCSA)*, pages 1–7.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov,

and Antonios Anastasopoulos. 2024. Dialect-bench: A nlp benchmark for dialects, varieties, and closely-related languages.

Songül Gündoğdu, Ergîn Öpengîn, Geoffrey Haig, and Erik Anonby, editors. 2019. *Current Issues in Kurdish Linguistics*, volume 1 of *Bamberg Studies in Kurdish Linguistics (BSKL)*. University of Bamberg Press.

Geoffrey Haig. 2001. Corpora of spoken and written varieties of kurdish.

Geoffrey Haig and Ergin Öpengin. 2018. Kurmanji kurdish in turkey: Structure, varieties, and status. *Linguistic minorities in Turkey and Turkic-speaking minorities of the periphery*, pages 157–229.

Geoffrey Haig, Donald Stilo, Mahîr Dogan, and Nils N. Schiborr. 2024. Wowa — word order in western asia: A spoken-language-based corpus for investigating areal effects in word order variation.

Hossein Hassani. 2021. Can linguistic distance help language classification? assessing hawrami-zaza and kurmanji-sorani.

Hanar Hoshyar Mustafa and Rebwar Nabi. 2022. Kurdish kurmanji dialect dataset for kurmanji lemmatization and spell-checker with spell-correction.

Abdullah Incekan and Geoffrey Haig. 2021. The corpus of contemporary written kurdish.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. 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*, pages 6282–6293, Online. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. Glotlid: Language identification for low-resource languages.

Aram Khaksar and Hossein Hassani. 2024. Shifting from endangerment to rebirth in the artificial intelligence age: An ensemble machine learning approach for hawrami text classification.

Hewa Salam Khalid. 2020. Kurdish language, its family and dialects. *International Journal of Kurdiname*.

Ben King and Steven Abney. 2013. Labeling the languages of words in mixed-language documents using weakly supervised methods. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1110–1119, Atlanta, Georgia. Association for Computational Linguistics.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang W. Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. 2024. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning.

Tessa Masis, Anissa Neal, Lisa Green, and Brendan O'Connor. 2022. Corpus-guided contrast sets for morphosyntactic feature detection in low-resource english varieties. *Proceedings of the first workshop on field linguistics*, pages 11–25.

Yaron Matras. 2019. Revisiting kurdish dialect geography: Findings from the manchester database. *Current Issues in Kurdish Linguistics. Bamberg: Bamberg University Press. 225-241.*

Dastan Maulud, Karwan Jacksi, and Ismael Ali. 2023. Towards a complete kurdish nlp pipeline: Challenges and opportunities. 17:1–17.

Peshmerge Morad, Sina Ahmadi, and Lorenzo Gatti. 2024. Part-of-Speech Tagging for Northern Kurdish. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 70–80, Torino, Italia. ELRA and ICCL.

Rudra Murthy, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2019. Addressing word-order di-

vergence in multilingual neural machine translation for extremely low resource languages. 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 3868–3873, Minneapolis, Minnesota. Association for Computational Linguistics.

Maria Nădejde, Siva Reddy, Rico Sennrich, Tomasz Dwojak, Marcin Junczys-Dowmunt, Philipp Koehn, and Alexandra Birch. 2017. Predicting target language ccg supertags improves neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 68–79, Copenhagen, Denmark. Association for Computational Linguistics.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9–16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Fatih Ozek, Bilgit Saglam, and Charlotte Gooskens. 2021. Mutual intelligibility of a kurmanji and a zazaki dialect spoken in the province of elazığ, turkey. *Applied Linguistics Review*.

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2025. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197.

Guilherme Penedo, Hynek Kydlíček, Loubna B. Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024a. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024b. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-Eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language.

Steven T. Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130.

Paul Rayson. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4):519–549.

Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2020. Understanding the effects of word-level linguistic annotations in under-resourced neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3938–3950, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nathan C Sanders. 2010. *A Statistical Method for Syntactic Dialectometry*. Doctor of philosophy, Indiana University.

Yves Scherrer. 2012. *Generating Swiss German Sentences from Standard German: A Multi-Dialectal Approach*. Ph.D. thesis, Université de Genève.

Christian Schuler and Raman Ahmad. 2023. Towards a complete mapping of kurdish dialectology [conference poster & abstract].

Thamar Solorio and Yang Liu. 2008. Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 973–981, Honolulu, Hawaii. Association for Computational Linguistics.

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2025. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset.

Givi Tavadze. 2019. Spreading of the kurdish language dialects and writing systems used in the middle east. 13(1).

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Mildred C. Templin. 1957. *Certain Language Skills in Children: Their Development and Interrelationships*, ned - new edition edition, volume 26. University of Minnesota Press.

Georgelle Thomas, R. Derald Hartley, and J. Peter Kincaid. 1975. Test-retest and inter-analyst reliability of the automated readability index, flesch reading ease score, and the fog count. *Journal of Reading Behavior*, 7(2):149–154.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Hadi Veisi, Mohammad MohammadAmini, and Hawre Hosseini. 2020. Toward kurdish language processing: Experiments in collecting and processing the asosoft text corpus. *Digital Scholarship in the Humanities*, 35(1):176–193.

Wikimedia. 2025. List of wikipedias by language group - meta.

Roy Xie, Orevaoghene Ahia, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. Extracting lexical features from dialects via interpretable dialect classifiers.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Omar F. Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial Evaluation Campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.

G. K. Zipf. 1935. *The Psycho-Biology of Language*. The Psycho-Biology of Language. Houghton, Mifflin, Oxford, England.

10. Language Resource References

Ahmadi, Sina and Q. Jaff, Daban and Ibn Alam, Md Mahfuz and Anastasopoulos, Antonios. 2024. *CORDI — Corpus of Dialogues in Central Kurdish*. ELRA Language Resources Association. PID https://github.com/sinaahmadi/CORDI.

Ahmadi, Sina and Sennrich, Rico and Karami, Erfan and Marani, Ako and Fekrazad, Parviz and Akbarzadeh Baghban, Gholamreza and Hadi, Hanah and Heidari, Semko and Dogan, Mahîr and Asadi, Pedram and Bashir, Dashne and Ghodrati, Mohammad Amin and Amini, Kourosh and Ashourinezhad, Zeynab and Baladi, Mana and Ezzati, Farshid and Ghasemifar, Alireza and Hosseinpour, Daryoush and Abbaszadeh, Behrooz and Hassanpour, Amin and Jalal Hamaamin, Bahaddin and Kamal Hama, Saya and Mousavi, Ardeshir and Nazir Hussein, Sarko and Nejadgholi, Isar and Ölmez, Mehmet and Osmanpour, Horam and Roshan Ramezani, Rashid and Sediq Aziz, Aryan and Salehi Sheikhalikelayeh, Ali and Yadegari, Mohammadreza and Yadegari, Kewyar and Zamani

Roodsari, Sedighe. 2025. *Parallel Corpora for Low-resourced Middle Eastern Languages*. Association for Computational Linguistics. PID https://github.com/DOLMA-NLP/PARME.

Haig, Geoffrey and Stilo, Donald and Dogan, Mahîr C. and Schiborr, Nils. 2024. *Word Order in Western Asia: A spoken-language-based corpus for investigating areal effects in word order variation.* PID https://fd-repo.uni-bamberg.de/records/gyws0-g4218.

Kurt, Fatih. 2020. *Kurdish Twitter Data*. PID https://github.com/ftkurt/kurdish-twitter-data.

Matras, Yaron. 2016. *The Dialects of Kurdish*. PID https://www.kratylos.org/ raphael/kurdish/database/.

Guilherme Penedo and Hynek Kydlíček and Vinko Sabolčec and Bettina Messmer and Negar Foroutan and Amir Hossein Kargaran and Colin Raffel and Martin Jaggi and Leandro Von Werra and Thomas Wolf. 2025. *FineWeb2*. PID https://huggingface.co/datasets/HuggingFaceFW/fineweb-2.