Kurdilex: Multigenre Corpus and Concordancer for Kurmanji Kurdish Language

Mahmoud Alhadji – Ma'moun Alshtaiwi

DOI: 10.18355/XL.2026.19.04.21

Abstract
The objective of this study is to create KurdiLex, a multigenre corpus and concordancer for the Kurmanji Kurdish language, to create an accessible environment to explore real language use across a variety of textual genres. The design and development processes were corpus-based. A searchable multigenre Kurmanji corpus was compiled and organized, and a concordancing system was developed to support Keyword in Context (KWIC) searching, word-frequency analysis, collocation analysis, bigrams, trigrams, n-gram analysis, and advanced search and filtering. For the display of the context of a searched keyword, KurdiLex also provides a long contextual display, e.g., the sentence before and the sentence after the sentence in which the searched keyword is found. KurdiLex enables users to search for and explore Kurmanji lexical items in authentic contexts, to determine their frequency and distribution, to study recurring collocations and to identify frequently occurring multi-word sequences. Extended context is useful for interpreting polysemous words, idiomatic expressions and context-dependent meanings. The combination of a multi-genre corpus with concordancing and complementary analytical functions enables the system to support empirical investigation of Kurmanji and contribute to research in Kurdish linguistics, lexicography, translation and computational linguistics.
Key words: Corpus-based linguistics, lexical patterning, contextualization, computational lexicography

1- Introduction

Kurmanji Kurdish is one of the main varieties of Kurdish and is spoken in several geographical regions. Kurmanji Kurdish is mostly written using the Kurmanji orthography, which is based on Latin. While interest in Kurdish NLP is growing, Kurmanji remains constrained by the availability of large-scale, accessible and linguistically diverse resources. Existing research has provided useful resources in building Kurdish corpora, dependency parsing, machine translation, dialect identification and other NLP tasks (Gökırmak & Tyers, 2017; Azzat, Jacksi & Ali, 2023). However, available resources are usually oriented toward specific computational applications, not full-scale, corpus-based lexical exploration.

The primary objective of KurdiLex (https://kurdilex.com) is to provide an integrated corpus environment for researchers, linguists, translators, language teachers, students and other users to investigate authentic Kurmanji language data in various genres. The multigenre orientation is particularly relevant, as linguistic patterns may vary according to the context of communication, the genre and the register. A corpus with different textual sources is therefore more likely to be representative of the use of Kurmanji than a corpus limited to one domain.

The multigenre corpus is designed to reflect the diversity of various textual domains, such as literature, newspapers, theatre, websites, and traditions. Combining several genres is a method to avoid limited representation of the language and offers researchers the possibility to investigate differences in vocabulary, style, register and phraseological patterns. It also allows for KWIC concordances, frequency analysis, collocations, prefixes, suffixes and n-grams to be used to compare lexical frequency and contextual use between genres.

The Concordancer is the main analytical tool in KurdiLex. It enables the user to search for a word or expression by means of a KWIC search and to view the surrounding language of

that word or expression. KurdiLex goes beyond the functionality of a conventional Concordancer by supplying additional information, i.e., the sentence before the word, the sentence with the word, and the sentence after the word. This information is required for analyzing polysemy, the contextual meaning of words, idiomatic expressions and discourse. The immediate context of a word is not always sufficient to answer these types of questions. KurdiLex packs in a bunch of handy tools for digging into text—like word-frequency checks, collocation analysis, n-gram searches, and some advanced filtering features. With these, you can start by spotting single words here and there, and then quickly shift to uncovering bigger patterns and common phrases. Frequency analysis shows you that words pop up the most, collocation analysis highlights that ones tend to team up, and n-gram analysis uncovers repeated word sequences. All of this makes it easier to see not just the words themselves, but how people actually use them together.

KurdiLex was born out of a clear need. Kurdish language tools and resources have come a long way, but there is still no all-in-one, Kurmanji-focused, multi-genre corpus with a built-in concordancer. KurdiLex steps in to fill that gap. It brings together real, varied Kurmanji texts and makes them easy to analyze with user-friendly corpus tools. This study walks through how KurdiLex was designed and built, highlights its main features, and looks at how it can push forward the study and computational analysis of Kurmanji Kurdish.

2- Literature Review

Corpus linguistics is the backbone behind KurdiLex. Instead of leaning on personal intuition or a few examples, it digs into real language data stored in electronic corpora. By analyzing huge piles of authentic texts, researchers can spot patterns that show up repeatedly. It is all about using computers to sift through the evidence and pull out what is really happening with the language (Biber *et al.*, 1998; Kennedy, 1998; McEnery & Wilson, 2001; McEnery & Hardie, 2012).

Sinclair (1991) points out that we should really pay attention to how words are used repeatedly, not just their meanings. The user of KurdiLex should take this idea and run with it. When someone enters a keyword, the tool grabs examples straight from the Kurdish corpus and shows exactly how the word appears in real sentences. This lets researchers look beyond whether the word pops up. They can see how often it shows up, which words hang around it, what phrases come out of it, and what kinds of situations it is actually used in.

This lets users see how often words show up across the whole corpus, giving a real sense of which words are common and which ones are not. With this information, it is easy to compare regular vocabulary with more unusual words. Frequency matters a lot when you're exploring which words stand out and what patterns keep coming up over and over (Kennedy, 1998; Biber *et al.*, 1998).

Later, Sinclair took the collocation further in 1991, showing just how crucial recurring word combinations are. With KurdiLex, you can dig into which Kurdish words tend to show up together. This unveils a lot about how words connect and what they really mean in context.

N-gram analysis takes KurdiLex a step further—it does not just look at single words. With N-grams, you can spot recurring sequences of two words or more. These patterns often highlight common expressions, technical terms, set phrases, or just the way people naturally string words together.

N-Gram analysis	Kurmanji Kurdish	English
Expression	*Biryar dan*	to make a decision
Technical term	*Zimanê xwezayî*	natural language
Phrase	*Parastina zimanê kurdî*	preservation of the Kurdish language

There is also an important difference between corpus-based and corpus-driven research. Tognini-Bonelli (2001) points out that some researchers start with specific questions and use corpus data to find answers, while others let the data itself reveal new patterns and insights. KurdiLex handles both angles. You can search for any Kurdish word or phrase you have in mind, but you can just as easily dive in and see what frequency, collocations, or recurring n-grams pop up organically.
Therefore, KurdiLex can be conceptualized as the computational implementation of core corpus-linguistic principles, particularly:

- **Concordancers and Their Applications: The KurdiLex Perspective**

In corpus linguistics, a concordancer is a tool that grabs every instance of a word or phrase from a text collection and shows them with their surrounding context (Sinclair, 1991; Hunston, 2002; McEnery & Hardie, 2012). With KWIC, the searched word is right in the middle, and you see snippets of text just before and after it. This lets researchers scan through a bunch of examples at once and catch patterns in how language is actually used. KurdiLex uses KWIC as one of its main features. If you search for a Kurdish word, KurdiLex pulls up all its appearances from the corpus and lays them out in context. You get to compare different cases and look deeper into how the word's meaning or usage changes. However, KurdiLex does not stop at traditional KWIC. It goes further by showing a wider chunk of context. Instead of just a short window around the word, you can see the sentence before, the sentence with the keyword, and the next sentence. That is especially useful for Kurdish, since a small KWIC window can leave you guessing about polysemous words, idioms, or meanings that depend on the broader conversation.
Take a word with multiple meanings—one KWIC line might not clear things up. With more of the surrounding sentences, KurdiLex helps you move from simply spotting the word to actually figuring out what it means in each context.
Concordancing also drives lexicographical work. Sinclair (1987) showed how crucial corpus evidence is for making dictionaries, and Atkins and Rundell (2008) stressed how it helps identify word senses, examples, grammar, and collocations. KurdiLex provides this kind of evidence, supporting Kurdish dictionary building and terminology research. It is not just for research either. Concordancers like KurdiLex are great for teaching. Data-driven learning has shown that students can dig into real language examples using concordances (Johns, 1991; Boulton, 2010). In Kurdish classrooms, KurdiLex could give both learners and teachers access to language data. Therefore, concordancing in KurdiLex is not just a search feature. It is the main link connecting users with authentic Kurdish language evidence.

- **Positioning KurdiLex**

Many corpus-analysis systems have shown the real value of computational concordancing. *Wordsmith Tools* (Scott, 2001), *AntConc* (Anthony, 2004), *Sketch Engine* (European

project), *CQPweb* (Hardie, 2012), and *#LancsBox* (Brezina, 2025) are the heavy hitters in this space—they have set the standard for what corpus tools can do. These systems focus on language analysis in a broad, language-independent way. KurdiLex, on the other hand, is not built for any language—it is designed specifically for Kurdish.

That matters because working with Kurdish brings a tangle of challenges. There are multiple dialects and different scripts, messy spelling variations, complex morphology, and not enough computational resources. The split between Sorani and Kurmanji only makes things harder. In addition, most Kurdish analysis tools are not particularly user-focused. KurdiLex is not just copying AntConc or Sketch Engine. Instead, it takes their best ideas and reshapes them for the realities of Kurdish data and user needs.

KurdiLex does not stand out because of sheer corpus size. Its real strength comes from several points working together: it focuses on Kurmanji, covers a range of genres, respects original sources, ensures everything is traceable, and brings together core analysis tools—like KWIC, frequency, collocation, and n-gram search-in one system. Plenty of Kurdish resources offer parts of this, but KurdiLex combines them in a way that gives it a distinct and much-needed role in Kurdish corpus research.

Resource	**Main Specialization**	**Possible Distinction of KurdiLex**
Pewan (Ahmadi et al: 2020; Ahmadi et al:2022)	Sorani News/Sorani press corpus, Sorani songs	KurdiLex is kurmanji multi-genre.
Kurmanji-MG (Gökırmak, & Tyers: 2017)	Syntactically annotated with a translated corpus (Short Sherlock Holmes Stroy translated by Segman (1944) and published in the Rohani journal in Damascus.	KurdiLex offers greater genre diversity and source documentation, although it is not yet syntactically annotated.
InterdialectCorpus (Ahmadi *et al*, 2022)	Parallel (Sorani, Kurmanji, Sorani - English, Kurmanji - English)/Statistical machine translation corpus	KurdiLex is primarily monolingual, with a focus on analysing Kurmanji in context
KTC (Abdulrahman *et al.* : 2019)	31 Sorani educational textbooks (K-12), 693,800 tokens.	KurdiLex focuses on Kurmanji and includes multiple text genres
Dasinya (Saeed & Jacksi, 2026).	A dataset and pre-processing pipeline for the Badini Kurdish dialect (subdialect of Kumanji)	KurdiLex focuses on Kurmanji and includes multiple text genres
AsoSoft (Veisi *et al*., 2020)	Large Sorani corpus. 190 M tokens.	KurdiLex focuses on Kurmanji and emphasizes source preservation and traceability.
KurCorpus 2B (Rawf, 2025)	Multidialectal (Sorani, Badini, Hawrami) corpus for model training	KurdiLex can prioritize fine-grained documentation, source traceability, and linguistic analysis rather than corpus size alone.

Folklore Corpus (Ahmadi *et al.*, 2020)	Sorani songs and folklore	KurdiLex incorporates folklore alongside a wider range of genres, providing broader textual coverage.

We can read the previous table in different dimensions:

1. **Language variety**: Many resources focus on Sorani variety. Our dedicated project focuses on Kurmanji in Latin alphabet characters.
2. **Genre coverage**: Kurdilex focuses on multigenre: poetry, novels, news, Web texts, and traditional songs.
3. **Corpus orientation**: the majority of databases are designed for a specific research or NLP purpose. In Kurdilex, we integrated resources for linguistics, lexicography, cultural heritage, and NLP.
4. **Original-Text Preservation**: in Kuridlex, original texts are preserved and linked to processed/normalized versions.
5. **Concordancing**: other projects are often available through selected paid platforms, but Kurdilex is an open-source web-based concordancer.
6. **Updating**: Some resources are relatively static. Kurdilex is a progressive, versioned addition of new documents.

KurdiLex stands out as a resource that adds something fresh to the Kurdish corpus landscape, not as a replacement for what is already out there. Where most existing corpora focus on one thing—like a single dialect (Sorani), text type, or a specialized NLP task—KurdiLex pulls together multiple threads. It zeroes in on Kurmanji written in the Latin script, covers a wide range of genres, and keeps the original texts intact and traceable. On top of that, it merges tools for both corpus linguistics and lexicographic work. In addition, KurdiLex preserves traditional songs, connects with a dedicated dictionary, and plans to document neologisms, all of which stretch its use beyond what you usually get from a corpus. In this way, KurdiLex is shaping up into an integrated platform—for language analysis, lexicon building, preserving cultural heritage, and supporting NLP—designed with Kurmanji at its core. The comparison can, therefore, be conceptualized as:
The distinctive contribution of KurdiLex therefore lies not necessarily in inventing a new corpus-analysis function, but in integrating established corpus-linguistic functions into a dedicated Kurdish concordancing environment.

- **Kurdish Language Resources and Corpora: The Need for KurdiLex**

Over the past decade, Kurdish computational resources have grown a lot. Compared to English, French, or German, Kurdish is far behind. Researchers have built corpora and developed machine translation, POS tagging, parsing, speech recognition, named-entity recognition, and other NLP tools (Hassani, 2018; Azzat, Jacksi & Ali, 2023). The Kurdish BLARK initiative stands out here; it is a focused effort to organize what language resources exist and make sense of what Kurdish language technology really needs (Hassani, 2018). Kurdish needs solid resources at every linguistic and computational level.
The *AsoSoft* corpus stands out as one of the key large-scale Kurdish text resources available today. Veisi *et al*. (2020) built and processed a significant Central Kurdish corpus, showing just how valuable it can be for computational linguistics research.
Abdulrahman *et al.* (2019) created the Kurdish Textbooks Corpus (KTC), which organizes educational texts into a structured dataset.
For KurdiLex, the developments mean we are no longer dealing with a total lack of Kurdish text data. The real issues now? Access, integration, exploration, and how to make this data genuinely usable for linguistic work.

A large corpus is not much use if researchers cannot search for words efficiently, see them in context, analyze frequency, find collocations, or dig into common n-grams. Without these tools, most of the corpus remains out of reach, especially for people who are not specialists. This is exactly where KurdiLex comes in. KurdiLex acts as a bridge—an analytical interface—between Kurdish corpora and the people who want to study them.

3- Methodology

- Corpus Data Collection and Organization

The first step focused on gathering real Kurdish-language texts and organizing them in a way that reflects the variety and history of the language. With KurdiLex, the goal was clear: bring together a broad sample of genres and periods so researchers can dig into Kurdish vocabulary in all sorts of contexts. The corpus splits the material into eight main categories: poetry, novels, theatre, newspapers, websites, specialized corpora, traditional song lyrics and dictionaries.

Domaine	Description
Poetry	The collection stretches all the way from the 16th century up to the 20th, giving a wide historical snapshot of how Kurdish words and expressions have changed.
Novels	Mostly from the 20th century.
Traditional songs lyrics	Drawn from the 20th century, add something special. They are not just full of expressive vocabulary—many of the terms and idioms in these songs do not even appear in standard Kurdish dictionaries. Including them means KurdiLex can capture vocabulary that usually slips past conventional linguistic resources.
Newspaper	Newspapers cover both the 20th and 21st centuries.
Specialized texts	Teaching textbooks.
Websites	For web content, the focus is on the present—recent articles, online features, and current writing.
Kurdish dictionaries	1 - English Kurdish Dictionary, Saladin Sadalah 2- *ferhenga kurdî Fatih Aydin* 3- *Erdal Ronahî* These dictionaries add another layer, especially in linking Kurdish words to English or to other foreign equivalents. KurdiLex was built so that people can keep adding to it—new languages,

- Data Cleaning and Organization

Once we gathered all the textual data, we rolled up our sleeves and got to work cleaning it up. We built a custom Python app to handle the whole process: cleaning and organizing the works. Since our texts often came as PDFs, we also used *AntFileCoverter* (2.1.0) to convert them into *.TXT* files we could actually use.

Cleaning is not just about tidying up; it is about cutting out technical noise while holding on to the real language inside these documents. We fixed encoding messes, took care of extra spaces and repeated lines, stripped out annoying page numbers, headers, footers, and those stubborn lines breaks; basically, all the little artifacts you run into when your data comes from all over the place.
With cleaner data in hand, we sorted everything by textual genre and attached all the relevant metadata. This makes searching the corpus much easier and lets us filter linguistic results.

- **Software Development and Technical Architecture**

We used Python not just for basic data processing and organization but also for building the main functions of the corpus. It handles massive amounts of Kurdish text and powers the algorithms behind the analysis. However, Python alone is not enough. We also relied

on several Python libraries—tools that make the whole process smoother. They handle text tokenization, searching, organizing the data, calculating frequencies, and other key tasks in corpus linguistics. Thanks to these libraries, we could automate repetitive parts and process huge volumes of Kurdish data quickly and efficiently. We also used the Flask framework to put the corpus online. With Flask, users get a simple web interface-they can search and interact with KurdiLex right from their browsers. Instead of just being a folder of files on a computer, KurdiLex works as a true online corpus thanks to this setup.

Future versions of the database will have the ability to add more texts, dictionaries, languages, metadata (for each document added: genre, source, time, and author), and analytical functions due to the flexible database architecture. The basic structure is similar but different from each other. Hence, KurdiLex is intended to be an evolving linguistic resource rather than a static corpus. Moreover, we intend to construct AI agents who can gather data from websites and newspapers to compare it with our existing data and identify inaccuracies.

Steps	Development Phase	Description	Technologies / Functions
1	Data Collection and Corpus Organization	Collection of Kurmanji textual resources, Classification and organization of texts	Text corpora, .TXT files, and Literature, poetry, traditional songs, websites, etc.
2	Cleaning and Normalization, Segmentation and Tokenization	Cleaning of text files and Segmentation of texts	Python, text processing
3	Corpus Indexing and Search Engine Development	Development of functions for querying the corpus and display of word occurrences	Python, Flask, search, genre filters and KWIC
4	Concordancer Development and Linguistic Analysis	Integration of tools for quantitative and contextual corpus analysis.	Frequency, collocations, N-grams
5	Web Interface Development and Data Export	Development of user interface for corpus searches and result visualization. Export of search results	HTML5, CSS3, JavaScript, Jinja2
6	Testing and Validation, Online Deployment	Verification of the search engine, Deployment of the KurdiLex platform for public web access.	Functional testing and validation

4- KurdiLex Processing Algorithm

The KurdiLex algorithm is a set of computational procedures that transform unrefined Kurdish texts into arranged and searchable collections. The general processing pipeline comprises:

The database contains the corpus, and a lexical index is generated that links each word to its occurrences and positions in the same corpora. This index enables the system to efficiently retrieve linguistic information without repeating across all available corpora.

The primary aim of KurdiLex is to locate words and expressions within their actual contexts of usage. It does not present the item as an isolated lexical item but draws its occurrences from original Kurdish texts and shows the surrounding context. This approach is particularly applicable to the study of meaning, collocations, phraseological patterns, lexical connections, semantic preferences, and contextual variation.
The primary search feature is based on Keyword in Context (KWIC) concordancing. KurdiLex retrieves occurrences from the corpus entered by a user and presents the keyword together with its contextual information within it.

- **Indexing and Frequency Analysis**

Tokenization is performed once at load time, after which KurdiLex builds a positional inverted index mapping each attested word form to the full list of its occurrences (document, line, token position).
Following tokenization, KurdiLex builds a positional inverted index that maps every word form attested in the corpus to the complete list of its occurrences, each recorded as a document identifier, a line number and a token position within that line. Separate indices are maintained for each genre-defined subcorpus, allowing queries to be restricted to a single text type. Contextual spans are not stored but reconstructed at query time from the pre-tokenized lines. This index provides the shared retrieval layer underlying both the search and the KWIC concordance functions.
KurdiLex also computes absolute frequency tables, both for the corpus as a whole and for each genre-defined subcorpus. These allow researchers to identify frequent word forms and to compare their distribution across sections of the corpus. Frequency counts are exported in tabular form, so that normalized measures—for instance, occurrences per million tokens—can be derived externally when subcorpora of unequal size are compared; such normalization is not currently performed by the system itself.

- **Data Preservation, Reproducibility, and Future Development**

KurdiLex adheres to a fundamental methodological principle of maintaining the original textual data. It is not recommended to overwrite the original text during cleaning or normalization. A log of transformations is essential for researchers to keep track of how the final corpus representation was created. This method enhances the corpus' reproducibility, transparency, and reliability. Future versions can enhance or revise normalization rules without compromising the original data. This is also possible. KurdiLex is designed to be a computational resource for Kurmanji and is extensible, dynamically flexible, and in the presence of an arbitrary corpus? Combining a range of historical and contemporary sources, along with traditional lexical material dictionaries.

5- Research Design

A corpus-linguistic approach is employed in this research using a specific methodology. The suitability of Design Research lies in its emphasis on creating, developing, implementing, and testing an innovative linguistic resource (KurdiLex) rather than on testing existing tools. The investigation employs a systematic approach of corpus formation, refinement and scrutiny to guarantee scientific precision and practical implementation. There are four distinct phases in the research:

- **Corpus Design**

Phase one involved defining the corpus' scope, identifying its genres, setting out criteria for selecting documents and designing the metadata schema. The corpus was designed to encompass a range of digital assets, including literary texts, newspapers, websites, novels, poetry, traditional Kurdish songs and specialized texts. The selection process took into

account language authenticity, document quality, legal availability, and genre representation.

- **Corpus Construction**

The second phase involved the collection and processing of data. The documents were gathered from various digital sources and converted into machine-readable text through direct extraction, OCR, or manual transcription, as needed. Text purging, Unicode normalization and clean up, metadata extraction (for duplicates), segmentation of sentences and tokenization were all part of the pre-processing process. In order to maintain a sense of linguistic authenticity, original texts were preserved alongside normalized versions, with each revision being recorded in synchronization through tracing the changes.

- **System Development**

The corpus was indexed using an inverted index structure to ensure efficient retrieval, including virtual keyboards and autocomplete. A web-based concordancer was created to facilitate KWIC searches, frequency analysis, genre filtering, and extended contextual retrieval. Its architecture also allows for the export of lexical statistics and frequency lists to be used later in linguistic analysis.

- **Evaluation**

Finally, both the corpus and retrieval system were evaluated. The quantitative evaluation included corpus statistics (documents, tokens and types, lexical diversity [clarification needed]], sentence counts and genre distribution), while the system evaluation evaluated retrieval accuracy, search response time, filtering performance, frequency correctness, and export functionality. This involved manually validating search results using a representative set of search queries.

6- System Features

The KurdiLex concordancer is a collection of corpus analysis tools designed for research on the Kurmanji Kurdish language, including lexical, grammatical and discourse-based aspects. These functions include retrieving lexical information and studying the contextual usage of words and phrases.

- **KWIC Search**

The Concordancer's core is dedicated to the KWIC search. It obtains all instances of a word or phrase specified by the user and exhibits them in its surrounding textual context. KurdiLex differs from conventional KWIC systems in that it presents an extended contextual display, including the preceding sentence, the sentence with the keyword, and what follows. A broader context supports understanding polysemous words (e.g., head; above; about with an abstract meaning), idiomatic expressions (example), and discourse relations without affecting the link between each instance or event and its original source document.

- **Word Frequency Analysis**

The system calculates the frequency of lexical items within specific subcorpora and across the entire corpus without human intervention. Using ranked frequency lists, users can identify frequent vocabulary, compare genres and lexical distributions, and examine terminology.

1	*bi*	20,251
2	*ji*	14,778
3	*li*	14,167
4	*xwe*	12,279

- **Collocation Analysis**

The collocation analysis module identifies words that frequently co-occur with a specific target lexical item within arbitrary context windows. This feature permits researchers to examine lexical connections, phraseological patterns, semantic preferences, and recurrent expressions.

Node	Collocation (bigram)	Frequency
dil	*dil bi*	120
dil	*ji dil*	91
dil	*bi dil*	71
dil	*min dil*	56

- **N-gram Analysis**

The system helps in the examination of formulaic language, idiomatic expressions, collocations and syntactic patterns by finding frequently occurring multiword expression (Altomonism) words. Language modelling and automatic text generation benefit from the use of n-gram statistics, which are also useful for natural language processing.

N	Most frequent n-gram	Frequency
2	*li ser* ("on, about")	2,479
3	*bi hev re* ("together, with each other")	316
4	*ji xwe re got* ("said to himself")	45

- **Advanced Search and Filtering**

The concordancer lets you run advanced searches where you can narrow things down by document details or language features. You can filter results by genre, author, publication year, source, corpus version, or choose between normalized and original text. It also handles exact word and phrase searches and wildcards, and lets you decide if you want case-sensitive results or not. These filters make it easier to find exactly what you need and dig into the corpus the way you want, especially when you're looking at specific types of documents.

- **Metadata-Based Retrieval and Export**

KurdiLex associates every document with structural metadata, its source file name, its genre category, and the subcorpus folder in which it is stored. Author and title information is encoded in the file naming convention. Every concordance line displays the source document and its genre, allowing results to be traced back to their origin, and search results, word lists, frequency statistics and n-gram tables can be exported as plain text or CSV files for further analysis.

Document	MetaData 1#Genre	MetaData 2# Subcorpus folder
Mehmed Uzun SIYA EVINE.txt	Novel (Roman)	Novels
Mem û Zîn - Kurdî-Ehmedê-Xanê.txt	Poetry (Helbest)	Poetry
EREB SEMO-ŞIVANÊ KURD.txt	Novel (Roman)	Novels

These fields can be used to restrict searches to a given genre or subcorpus, and frequency lists, lexical inventories and concordance results can be exported as TXT or CSV files. This supports reproducibility and facilitates further linguistic and statistical analysis outside the interface.

7- Evaluation

We will look at KurdiLex from four different angles: how well the system performs, how accurate and efficient it is, how easy it is to use, and how it stacks up against other concordancers out there. The goal is simple: find out if KurdiLex really gives people reliable, fast, and easy-to-use tools for analysing the Kurmanji Kurdish corpus.

- **Performance Evaluation**

We put KurdiLex to the test by seeing how fast it responds to different corpus sizes. We tried out a few subsets: a small set with about 2,075 words, a medium one with a million words, and a large set that packs in 2 million. For each of these, we timed how long it takes to do the big stuff: KWIC searches, checking word frequencies, running collocation analysis, extracting n-grams, and handling more searches that are advanced. We looked at how quick the searches are, how long it takes to extract n-grams or analyse collocations, how many tokens the system can process each second, how much memory it uses, and how stable everything stays—especially with the big data.

- **Accuracy and Efficiency**

To check how accurate KurdiLex is, we compared the system's automatic results with manually verified data from a good chunk of the Kurmanji corpus. We picked out things like KWIC results, collocations, and n-grams, and went through them by hand to see if KurdiLex really catches the right words and the connections between them.
For the n-grams, we looked closely at whether the system finds real sequences of words and gets their frequencies right. When it comes to collocations, we want to see if it spots words that actually pop up together with a certain word more often than chance would suggest.

- **Usability Testing**

We also planned to run a usability study for KurdiLex, inviting folks who actually use it—Kurdish language researchers, linguists, teachers, translators, and students. We will ask them to complete typical tasks: searching for a word, checking KWIC concordance lines, spotting collocations, making n-grams, and using advanced filters.
They judged how easy it is to learn and navigate the system, whether the interface makes sense, how pain-free it is to search and find results, how useful the analysis features are, and how satisfied they feel overall.
We handed out a five-point Likert scale questionnaire, with options from "strongly disagree" to "strongly agree." Moreover, we will keep track of the task-completion rates and how long each task takes so we can actually measure how well the system works in numbers.

- **Comparison with Existing Concordancers**

KurdiLex is set to go head-to-head with well-known corpus-concordancing tools, with a close look at how they actually work for Kurmanji Kurdish. Think of software like*AntFileCoverter* (2.1.0), but tweaked or compatible with Kurdish—that is the level of comparison. We are not just counting how many features each tool offers. What really matters is how well those features serve Kurmanji Kurdish. So, things like Kurdish spelling differences, the language's tricky morphology, how tools break down words (tokenization), how flexible the search functions are, and whether they offer Kurdish-specific resources—all of that gets careful attention.

8- Results and Discussion

KurdiLex has helped build an online corpus and concordancer specifically for the Kurmanji Kurdish language. It has used a combination of heterogeneous texts from the Kurdish language and offers features for lexical searches, context analysis, frequency calculation, and corpus mining. The findings show that the proposed architecture has helped build a linguistic database from diverse Kurdish textual sources.
The result is a corpus of various textual genres, such as poetry, novels, newspapers and websites, theatre productions, traditional song lyrics, specialized texts, and dictionaries. This variety provides a more comprehensive portrayal of the use of Kurdish throughout history, culture and literature.
The corpus is enriched by historical poetry, which not only extends the study of contemporary Kurdish but also allows for the examination of changes in vocabulary and language. Novels and newspapers are also a source of literary and journalistic language from the twentieth and twenty-first centuries. Newer websites offer fresh words and phrases, especially those related to current affairs and online communication.
Additionally, traditional songs are a significant aspect. Culturally embedded lexicon and phrases that are not commonly found in modern written texts or existing dictionaries are also included to enrich the corpus. These examples are noteworthy. The traditional-song component is a crucial complementary resource for linguistic research.
This allowed Python to automatically clean, organize, and sort vast amounts of textual data, normalize it, and prepare it.
One of the key results from this pre-processing process is that it preserves a relationship between the original texts and their processed versions. Instead of treating cleaning and normalization as "impossible" to change, KurdiLex is focused on maintaining the original linguistic forms and documenting important changes. Authentic texts can contain orthographic variations, Unicode inconsistencies, and multilingual elements, making this an important approach to dealing with Kurdish.

The key benefit of using KurdiLex is its ability to retrieve lexical items from the contexts in which they are used. The search and KWIC functionality allow researchers to examine the word in question as it occurs before and after its answer.
The advantage over traditional dictionaries is that they usually exhibit the same lexical item without any reference to its actual textual environment. This feature is particularly useful in this context. By using KurdiLex, researchers can examine the usage of a word in everyday language and observe its repeating linguistic patterns.
The exploration of a term, such as *azadî*, can yield several instances from various genres and sources.
Collocations, phraseological patterns, semantic preferences, lexical associations, and contextual meanings can be studied with the help of the KWIC functionality. Researchers can compare different contexts where the same lexical form takes place, making it an important factor in studying polysemous words.
Frequency analysis is another significant outcome. With KurdiLex, it is possible to determine how frequently lexical items are present in the corpus.
Identifying frequent and less frequently used words is possible through empirical evidence obtained from frequency information. Comparing lexical distributions across different genres/eras can be made easier. The occurrence of a specific word in newspapers or the frequency of that item in poetry can be investigated by examining historical data. Additionally, some researchers have found evidence supporting this observation.
KurdiLex incorporates dictionaries, which go beyond the traditional corpus concordancer. Supplementary lexical information is available in dictionaries, which can also provide foreign-language equivalents for languages such as English.
This element provides an opening for the future development of KurdiLex as a multilingual resource.' Future versions of the database will have the option to incorporate additional languages, such as French and Arabic.
The extension would be particularly beneficial for teaching, translating, defining terminology, creating glossaries, and conducting computational linguistics in multiple languages.
The outcomes indicate that KurdiLex offers a cohesive platform for conducting research in Kurdish corpus. Its integration of various textual sources, automation of pre-processing, lexical indexing and search, as well the use of KWIC concordance and frequency analysis, allow for an empirical study of Kurdish vocabulary.
A significant contribution is the integration of traditional and possibly less emphasized vocabulary. The inclusion of lyrics and other important sources from around the world can enable KurdiLex to create searchable vocabulary that may not be found in existing dictionaries or modern written archives.

9- Conclusion

KurdiLex findings suggest that it has the potential to create a computational model for conducting searches and analysis of Kurmanji Kurdish data. The key benefit is the integration of various textual sources, along with computational processing and contextual lexical retrieval. Through this system, it is evident that a Kurdish corpus can be constructed as an interactive research tool that links lexical forms to their authentic contexts, sources and frequencies rather than just being arranged in digitized forms. This provides a solid foundation for future research in Kurdish linguistics, lexicography, terminology, translation, language teaching, and computational linguistics.
The linguistic rules that are available for Kurmanji play a crucial role in determining the quality of automatic cleaning and normalization. It may be necessary to manually confirm ambiguous or historically inconsistent forms. By including additional documents, the corpus can be expanded to increase its representativeness and facilitate more dependable statistical comparisons.
Further areas of focus for the future are linguistic annotation, which will include more complex methods such as part-of-speech tagging, lemmatization, morphological analysis,

named-entity recognition and semantic annotation. These functions would facilitate researchers in pursuing more sophisticated linguistic analysis.

Bibliographic references

Abdulrahman, R. O., Hassani, H., & Ahmadi, S. (2019). Developing a fine-grained corpus for a less-resourced language: The case of Kurdish. arXiv preprint arXiv:1909.11467.

Ahmadi, S., Hassani, H., & Abedi, K. (2020a, May). A corpus of the Sorani Kurdish folkloric lyrics. In Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU), Collaboration, and Computing for Under-Resourced Languages (CCURL) (pp. 330–335).

Ahmadi, S., Hassani, H., & Jaff, D. Q. (2022). Leveraging multilingual news websites for building a Kurdish parallel corpus. Transactions on Asian and Low-Resource Language Information Processing, 21(5), 1–11.

Anthony, L. (2005, March). AntConc: A learner- and classroom-friendly, multi-platform corpus analysis toolkit. In IWLeL 2004: An Interactive Workshop on Language E-Learning (pp. 7–13). Waseda University.

Atkins, B. S., & Rundell, M. (2008). The Oxford guide to practical lexicography. Oxford University Press.

Azzat, M., Jacksi, K., & Ali, I. (2023). The Kurdish language corpus: State of the art. Science Journal of University of Zakho, 11(1).

Biber, D., Conrad, S., & Reppen, R. (1998). Corpus linguistics: Investigating language structure and use. Cambridge University Press.

Boulton, A. (2010). Learning outcomes from corpus consultation. In Exploring New Paths in Language Pedagogy: Lexis and Corpus-Based Language Teaching (pp. 129–144).

Brezina, V. (2025). Corpus linguistics and AI: LancsBox X in the context of emerging technologies. International Journal of Language Studies, 19(2), 75.

Esmaili, K. S., Eliassi, D., Salavati, S., Aliabadi, P., Mohammadi, A., Yosefi, S., & Hakimi, S. (2013). Building a test collection for Sorani Kurdish. In 2013 ACS International Conference on Computer Systems and Applications (AICCSA) 1–7. IEEE.

Gokirmak, M., & Tyers, F. (2017). A dependency treebank for Kurmanji Kurdish. In Proceedings of the Fourth International Conference on Dependency Linguistics (pp. 64–72).

Hardie, A. (2012). CQPweb—combining power, flexibility and usability in a corpus analysis tool. International Journal of Corpus Linguistics, 17(3), 380–409.

Hassani, H. (2018). BLARK for multi-dialect languages: Towards the Kurdish BLARK. Language Resources and Evaluation, 52(2), 625–644.

Hunston, S. (2002). Corpora in applied linguistics. Cambridge University Press.

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials (Vol. 4, pp. 1–16).

McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press.

McEnery, T., & Wilson, A. (2001). Corpus linguistics (Vol. 256). Edinburgh University Press.

Mrkva, J., Mrkvova, A., & Kilmanova, F. (2021). Impact on learning French. Contemporary Education, 12(2).

Rawf, K. M., Abdullah, A., Hussein, A., & Mohammed, H. (2025). KurCorpus 2B: A multidialectal 2-billion-token corpus for Kurdish language modeling (Version 1). Mendeley Data.

Saeed, V. A., & Jacksi, K. (2026). Dasinya: A dataset and preprocessing pipeline for the Badini Kurdish dialect. Data in Brief, 113, 129.

Scott, M. (2001). Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith Tools suite of computer programs. In Small Corpus Studies and ELT 47–67.

Sinclair, J. (1987). Looking up: An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary.
Sinclair, J. (1991). Corpus, concordance, collocation. Oxford University Press.
Tognini-Bonelli, E. (2001). Corpus linguistics at work (Vol. 6). John Benjamins.
Thompson, P. (2014). Corpus linguistics: Method, theory and practice. [Review of the book Corpus Linguistics: Method, Theory and Practice, by T. McEnery & A. Hardie].
Veisi, H., MohammadAmini, M., & Hosseini, H. (2020). Toward Kurdish language processing: Experiments in collecting and processing the AsoSoft text corpus. Digital Scholarship in the Humanities, 35(1), 176–193.

Words: 6725
Characters: 45 528 (25 standard pages)

Mahmoud Alhadji
University of Quebec in Outaouais (UQO)
Quebec
Canada
ORCID: 0009-0007-9860-4003
alhaj.mahmoud001@gmail.com

Ma'moun Alshtaiwi, PhD
Faculty of Arts
Department of Modern Languages Yarmouk University
Irbid
Jordan
ORCID: 0000-0003-4458-7694
Scopus ID: 57223021738
m.alshtaiwi@yu.edu.jo