Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Qaamuuska-NLP: A Structured 46K-Entry Somali Lexicon Extracted from a Print Dictionary

Abdullahi Mohamud Ahmed
Ogaal Labs
Abdulkadir Ugas
Ogaal Labs
Abdirashid Omar
rashiidmatan@gmail.com

Ogaal Labs
Abdulhakim Ismail
Ogaal Labs

Research Article

Keywords: Somali NLP, computational lexicography, lexical resources, dictionary digitization, morphology, low-resource languages

Posted Date: September 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10945954/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Qaamuuska-NLP: A Structured 46K-Entry Somali Lexicon Extracted from a Print Dictionary

Preprint, August 2026

Abdullahi Mohamud Ahmed[1], Abdulkadir Ugas[1,2], Abdirashid Omar[1], and Abdulhakim Ismail[1]

[1] Ogaal Labs (info@ogaallabs.com)

[2] Department of Bigdata, Chungbuk National University (ugas@chungbuk.ac.kr)

Abstract

Somali NLP still has relatively few structured lexical resources that preserve the linguistic information found in traditional dictionaries. We present **Qaamuuska-NLP**, a machine-readable reconstruction of *Qaamuuska Af-Soomaaliga*, the 2012 monolingual Somali dictionary by Puglielli and Mansuur. The resource contains **46,314 dictionary records** extracted directly from the PDF's native text layer using a deterministic, rule-based pipeline, without OCR or machine-learning-based extraction. Where available in the source, records include part of speech, noun gender, verb class and transitivity, plural information, subject domains, synonym references, cross-references, homonym indices, and numbered definitions. The original entry text is retained alongside the parsed representation to support inspection against the source.

We describe the extraction pipeline, the structure and coverage of the resulting resource, and the main classes of residual parsing errors. We also examine the distribution of grammatical and lexical information in the dictionary and use a simple surface-form prediction probe to test how strongly selected morphological labels are reflected in Somali word forms. Qaamuuska-NLP complements existing Somali morphological, lemmatization, corpus, and multi-dictionary lexical resources by preserving the structure of a single major monolingual dictionary in a reproducible computational form. The extraction code and aggregate statistics are being prepared for release; public redistribution of the complete structured lexical dataset remains subject to rightsholder permission.

***K*eywords** Somali NLP · computational lexicography · lexical resources · dictionary digitization · morphology · low-resource languages

1 Introduction

Somali remains comparatively under-resourced in natural language processing, particularly when compared with languages for which large lexical databases, annotated corpora, morphological tools, and machine-readable dictionaries are widely available. Recent work has expanded the Somali NLP ecosystem through information-retrieval corpora, language models, lemmatization resources, morphological tools, and other task-specific datasets [1–3]. At the same time, Somali has a much longer tradition of lexicographic work, including major monolingual and bilingual dictionaries containing grammatical and semantic information that is difficult to recover from running-text corpora alone.

One of the most important of these resources is *Qaamuuska Af-Soomaaliga* [4], a large monolingual Somali dictionary developed through a long-running Somali lexical project. Its entries contain more than headwords and definitions: the dictionary records grammatical categories, noun gender, verbal information, plural forms, semantic distinctions, subject labels, synonyms, homonyms, and cross-references. This makes it valuable not only as a reference work for human readers, but also as a potential source of structured lexical information for computational research.

The published dictionary, however, is distributed as a fixed two-column PDF rather than as a documented fielded dataset. Although its text can be extracted digitally, the lexical structure is still expressed through typography, abbreviations, entry ordering, superscript markers, punctuation, and recurring dictionary conventions. A plain text extraction therefore does not directly provide records that can be searched by grammatical field, linked through cross-references, or used systematically in NLP pipelines.

This paper presents **Qaamuuska-NLP**, a structured reconstruction of the dictionary containing **46,314 extracted records**. Rather than applying OCR to rendered pages, we work directly from the PDF's native text layer. The source has a sufficiently regular layout and entry structure to support a deterministic extraction pipeline: the two page columns are read separately, entry boundaries are identified from recurring headword and part-of-speech patterns, and dictionary-specific conventions are mapped into explicit fields. The pipeline does not use a trained extraction model or a generative language model.

Each resulting record retains its headword and available grammatical and lexical information, including part of speech, gender, verb class and transitivity, plural metadata, subject-domain labels, synonym references, numbered definitions, homonym indices, and dictionary cross-references. We also preserve the unparsed source entry text for each record. This allows the structured representation to remain connected to the dictionary material from which it was produced rather than replacing it with an opaque normalized representation.

Qaamuuska-NLP is situated within an existing history of Somali computational lexicography. Earlier projects have developed electronic dictionary systems, multi-dictionary lexical databases, finite-state morphological resources, corpus-linked lexicons, and more recent lemmatization datasets. In particular, the RCF-

SC/somMorph work combined several Somali dictionaries—including *Qaamuuska Af-Soomaaliga*—into a large lexical and morphological resource [5]. Our aim is therefore not to claim that structured Somali lexical resources did not previously exist. Instead, Qaamuuska-NLP focuses on a different resource design: a **single-source reconstruction of one named monolingual dictionary**, with the extraction procedure and its resulting structure documented directly.

This distinction matters because different Somali lexical resources represent different units and serve different purposes. A dictionary record is not equivalent to a morphological root, generated inflected form, corpus token, or normalized lemma. Preserving a dictionary as a structured resource also retains information that is often outside the scope of morphology-oriented datasets, including sense definitions, editorial cross-references, homonym distinctions, domain labels, and source-specific grammatical information. Qaamuuska-NLP is intended to complement, rather than replace, existing Somali resources.

We characterize both the resource and the extraction process. The resulting dataset contains 46,314 records, including 25,240 records with at least one definition and 21,032 cross-reference entries. Nouns and verbs account for the overwhelming majority of the dictionary, and the extracted structure captures noun gender, verbal conjugation class and transitivity, synonym links, plural metadata, homonym markers, and subject-domain annotations where these are present in the source. We report internal consistency diagnostics, unresolved references, duplicate-key cases, and residual parsing artifacts, together with a manually reviewed sample of extracted records.

We also examine whether selected grammatical labels in the dictionary show systematic relationships with the surface form of Somali words. Using a simple suffix-based prediction probe with a headword-disjoint split, we test noun gender, verb conjugation class, and coarse part of speech. This experiment is intended as a diagnostic of the lexical information contained in the resource rather than as an evaluation of a full Somali morphological analyzer.

The main contributions of this work are:

- **A structured reconstruction of *Qaamuuska Af-Soomaaliga*** containing 46,314 dictionary records and preserving a broad range of grammatical, lexical, semantic, and cross-reference information from the source.
- **A deterministic and reproducible extraction pipeline** designed for the dictionary's native two-column PDF text layer, avoiding both OCR and learned information-extraction models.
- **A detailed characterization of the resulting lexical resource**, including part-of-speech distributions, morphology, definitions, homonymy, cross-references, synonym links, and subject-domain information.
- **An analysis of extraction quality and morphological label predictability**, combining internal consistency checks, source comparison, error reporting, and a simple surface-form prediction probe.
- **A reproducibility package in preparation** comprising the extraction code, schema, aggregate statistics, validation materials, and an interactive research dashboard. Public redistribution of the complete lexical content remains subject to rightsholder permission.

The remainder of the paper is organized as follows. Section 2 reviews prior Somali lexical resources, dictionary digitization methods, and interoperability standards. Section 3 describes the source dictionary, extraction pipeline, record structure, parsing challenges, character handling, and quality checks. Section 4 presents a descriptive analysis of the extracted lexicon. Section 5 discusses potential NLP uses of the resource, Section 6 presents the surface-form prediction analysis, and Section 7 concludes with limitations and directions for future work.

2 Related Work

Qaamuuska-NLP sits at the intersection of Somali computational lexicography, dictionary digitization, and lexical-resource interoperability. Somali has a longer history of computational lexical work than is sometimes reflected in recent NLP literature, including dictionary databases, morphological analyzers, corpus-linked lexicons, and more recent lemmatization resources. Our work builds on this history but addresses a narrower task: reconstructing a single major Somali monolingual dictionary as a structured, reproducible computational resource.

2.1 Somali Lexical and NLP Resources

Computational work on Somali lexicography predates most recent NLP datasets. The Somali Language Lexical Project, described by Puglielli [6], began from earlier Somali dictionaries and textual sources and produced the lexical foundation for a series of bilingual and monolingual dictionary projects. The 1985 Somali–Italian dictionary was subsequently processed computationally to support inverse lookup, grammatical-category lists, lexical querying, and further development of the Somali Monolingual Dictionary. The same project also described work toward an automatic morphological analyzer and a lexical database linked to the developing monolingual dictionary. These efforts are important predecessors to current Somali NLP resources, even though the underlying database of the later monolingual dictionary does not appear to have been released as a standalone public machine-readable resource.

The closest large-scale computational resource to our work is the Redsea Cultural Foundation Somali Corpus and its associated lexical database [5]. That project digitized and normalized material from several Somali dictionaries, including *Qaamuuska Af-Soomaaliga* [4], and combined them into a shared lexical resource. The reported database contained 47,020 base lemmas with basic part-of-speech information, while the accompanying somMorph system generated approximately 1.2 million morphologically tagged forms. The resource also supported dictionary lookup, spelling correction, corpus search, POS tagging, and word profiles containing information such as definitions, spelling variants, synonyms, antonyms, and corpus examples. Because this resource aggregates several dictionaries and corpus-derived information, its counts and structure are not directly comparable to those of a single-dictionary extraction.

Earlier work also demonstrated that Somali print dictionaries could be converted into structured digital applications.

Thieberger and Faragaab [7] converted a digital version of Zorc and Osman's Somali-English Dictionary into an application-oriented structured format containing approximately 26,000 words and definitions, with fields such as headword and part of speech explicitly marked. More recently, Jibril and Badel [8] extracted 1,592 Somali homographs and their definitions from *Qaamuuska Af-Soomaaliga* for homograph-disambiguation research. These efforts show that Somali dictionary material has previously been digitized for computational use, although with different source dictionaries, coverage, and research objectives.

A separate line of work has focused on lemmatization and morphological normalization. Mohamed and Mohamed [9] introduced a rule-based Somali lemmatization approach supported by a lexicon of 1,247 roots and 7,173 derived forms. Gedi et al. [10] subsequently developed a substantially larger purpose-built lemmatization resource containing 5,584 roots and 78,629 POS-tagged derivatives, together with a web-based crowdsourcing and validation platform. Other machine-readable resources, including the LORELEI Somali language pack [11], GiellaLT's finite-state Somali analyzers, and corpus resources such as HaBiT/soWaC [12], provide additional morphological, lexical, or corpus-level information.

These resources differ substantially in their units of representation. A dictionary record, a lemma, a morphological root, a generated derivative, and a corpus word form are not equivalent counting units. Qaamuuska-NLP therefore does not position its 46,314 records as directly larger than these resources. Instead, it complements them by preserving the structure of one named monolingual dictionary and representing its grammatical, morphological, definitional, and cross-reference information in a single structured dataset.

2.2 *Digitizing Legacy Dictionaries*

Converting legacy dictionaries into machine-readable lexical resources is an established problem in computational lexicography [13–15]. The task involves more than recovering text: dictionary entries encode information through layout, typography, abbreviations, punctuation, indentation, and recurring field order. Different digitization pipelines therefore separate text recovery from the reconstruction of lexical structure.

Rule-based approaches are particularly useful when a dictionary has stable and predictable conventions. Maxwell and Bills [16] describe a workflow in which dictionary entries are recovered from digitized pages and then parsed using deterministic structural rules and finite-state methods. Their work highlights the importance of treating page layout, entry boundaries, and internal dictionary structure as separate stages rather than assuming that plain extracted text already reflects the intended lexical organization.

Learned approaches have also been used for dictionary structuring. Khemakhem et al. [17] introduced GROBID-Dictionaries, which uses cascading Conditional Random Field models to identify document regions, lexical entries, and internal structures before serializing them in TEI. Bowers et al. [18] applied this approach to a historical Mixtec dictionary, illustrating its use in the digitization of lexical resources for an under-resourced language. More recent work has explored large language models for dictionary information extraction. Jumashev et al. [19], for example, use language models to map inconsistently formatted Russian–Kyrgyz dictionary text into structured JSON records, while recent multilingual dictionary-digitization frameworks similarly separate text recognition from lexicographic parsing [20].

Our setting differs from OCR-centered dictionary digitization [21, 22] in an important way. *Qaamuuska Af-Soomaaliga* is distributed as a two-column PDF with an extractable native text layer. We therefore do not need to reconstruct characters from page images. Instead, the main problems are reading the columns in the correct order, identifying entry boundaries, and mapping the dictionary's recurring grammatical and typographic conventions into explicit fields. Given the regularity of this particular source, we use deterministic extraction rules rather than a learned sequence-labeling or generative model. The aim is not to propose a general parser for arbitrary dictionaries, but to provide a transparent and reproducible extraction procedure for this source.

2.3 *Lexical Representation and Interoperability*

A structured dictionary becomes more reusable when its fields can be mapped to established lexical and linguistic standards. TEI Lex-0 provides a constrained TEI-based model for representing common dictionary structures such as forms, grammatical information, senses, definitions, and cross-references [23]. OntoLex-Lemon provides a complementary linked-data model for lexical resources and their relationships to concepts and external knowledge representations [24]. These frameworks serve different purposes: TEI Lex-0 is closely aligned with the documentary structure of dictionaries, whereas OntoLex-Lemon is designed for interoperable lexical data on the Semantic Web.

For grammatical annotation, Universal Dependencies provides a widely used cross-linguistic inventory of part-of-speech categories and morphological features [25]. Our current version preserves the source dictionary's own grammatical codes and normalized internal categories rather than claiming direct compliance with these standards. Mapping the resource to Universal Dependencies and producing TEI Lex-0 or OntoLex-Lemon representations are therefore interoperability targets for future versions rather than features of the present extraction.

Taken together, prior work shows that Somali already has computational dictionaries, multi-source lexical databases, morphological analyzers, corpus lexicons, and lemmatization resources. Qaamuuska-NLP does not claim priority over that broader history. Its contribution is more specific: it provides a reproducible, single-source structured extraction of *Qaamuuska Af-Soomaaliga* [4], retaining 46,314 dictionary records and the grammatical, morphological, definitional, and reference information encoded in the source where those fields are present.

3 Methodology

We construct Qaamuuska-NLP by converting the dictionary body of *Qaamuuska Af-Soomaaliga* into structured records with a deterministic Python pipeline. The source PDF contains an extractable native text layer, so the pipeline does not use OCR. Instead, it addresses the main structural problems introduced by the two-column page layout, dictionary-specific abbrevia-

tions, superscript homonym markers, and recurring grammatical notation.

The extraction is intentionally source-specific. Our goal is not to build a general parser for arbitrary dictionaries, but to recover the structure of this particular dictionary in a transparent and reproducible way.

3.1 Source Dictionary

The source is *Qaamuuska Af-Soomaaliga* (Dictionary of the Somali Language), edited by Annarita Puglielli and Cabdalla Cumar Mansuur and published by Roma Tre Press in 2012 [4]. It is a major monolingual reference dictionary for Somali and contains grammatical as well as definitional information for its entries.

We process the main dictionary body, corresponding to PDF pages 24–916. The pages use a regular two-column layout, and the PDF provides a usable text layer. Text is therefore extracted directly with `pdfplumber` rather than reconstructed from rendered page images.

This distinction is important for the extraction design. OCR-oriented pipelines must first recover characters before recovering lexical structure. In our case, character recognition is not the primary problem; the main task is to reconstruct the intended reading order and identify the dictionary conventions that separate one record and one field from another.

3.2 Extraction Pipeline

The pipeline is implemented in Python. Its external dependencies are limited to `pdfplumber` for PDF text extraction and `tqdm` for progress reporting, while regular-expression parsing and JSON serialization use Python's standard library. Parsing is performed in Unicode mode and does not rely on a trained classifier or generative model.

The extraction proceeds in four stages.

3.2.1 Stage 1: Column-aware text extraction

Each dictionary page contains two text columns. Extracting the page as a single text block can mix lines from the left and right columns and corrupt the intended reading order.

We therefore divide each page at its horizontal midpoint and extract the two columns independently. The left-column text is processed before the right-column text, reproducing the order in which the dictionary is intended to be read. Short running headers at the top of the columns are removed using source-specific heuristics.

The resulting text is appended to a continuous buffer so that entries can be recovered even when their content extends across page boundaries.

3.2.2 Stage 2: Entry segmentation

Dictionary records are located using a recurring entry-start pattern. An entry normally begins with a headword, may include a superscript homonym marker, and is followed by one of the dictionary's recognized part-of-speech codes.

The parser scans the extracted text for this entry-start signature. The span between one recognized entry start and the next is treated as the body of a candidate dictionary record.

This approach relies on the regularity of the source rather than on a learned sequence model. The entry-start rules are therefore specific to the conventions used by *Qaamuuska Af-Soomaaliga*.

3.2.3 Stage 3: Field parsing

Each candidate record is then decomposed into structured fields.

The parser extracts, where present:

- the headword;
- the homonym index;
- the original part-of-speech code;
- a normalized internal part-of-speech category;
- noun gender;
- verb conjugation class;
- verb transitivity;
- noun plural information;
- parenthetical verbal information;
- subject-domain labels;
- numbered definitions;
- synonym references; and
- dictionary cross-references.

Numbered definition sections are represented separately rather than stored as one undifferentiated definition string.

Not every dictionary record contains every field. The resulting schema therefore reflects the information explicitly available in the source rather than forcing a uniform set of annotations onto all records.

3.2.4 Stage 4: Cross-reference resolution

A substantial portion of the dictionary consists of records that redirect the reader to another headword rather than supplying an independent definition. These are signaled in the source by dictionary-specific markers such as `ld` and `eeg`.

After the initial extraction, a second pass builds an index over the extracted headwords and attempts to connect each redirect record to the integer identifier of its target record. Where the source reference contains a homonym marker, that information is used during resolution.

Of the **21,032 redirect records**, **20,075** are linked successfully to an extracted target, corresponding to a resolution rate of **95.4%**. The remaining **957 records (4.6%)** are retained as unresolved references rather than discarded.

The extraction run is checkpointed every 50 pages and can be resumed after interruption. A unit-test suite covering representative dictionary patterns is used to guard the parser against regressions when extraction rules are modified.

3.3 Record Structure

The canonical extraction output is JSON, with one structured object for each dictionary record. Each record receives a sequential integer identifier during extraction.

A record includes the parsed fields available for that entry together with `raw_body`, which stores the original extracted entry text before field parsing. Preserving this source text is important because it allows the structured representation to be inspected against the material from which it was produced and prevents information outside the current schema from being silently discarded.

The integer `id` is a surrogate identifier rather than a linguistic key. The combination of headword and homonym index is useful for lookup, but it is not guaranteed to be unique in the extracted data and is therefore not used as the sole record identifier.

The current schema is designed around the structure of the source dictionary rather than an external lexical standard. Mapping the fields to interoperability formats such as TEI Lex-0 or OntoLex-Lemon is left to future work.

Table 1 summarizes the coverage of the principal extracted fields over the full collection.

Table 1: Coverage of principal extracted fields.

Field / property	Records	% of all records
Total records	46,314	100.0
Definition(s)	25,240	54.5
Redirect (`ld` / `eeg`)	21,032	45.4
Redirect resolved to target ID	20,075	43.3
Gender annotation	34,570	74.6
Verb-class annotation	11,444	24.7
Synonym reference(s)	11,415	24.6
Homonym index	9,212	19.9
Noun plural metadata	4,698	10.1
Subject-domain label	1,945	4.2
Predicative-only (`khabar`)	529	1.1

These percentages use the full set of 46,314 records as the denominator. Section 4 additionally discusses fields relative to the grammatical categories for which they are relevant.

3.4 Source-Specific Parsing Challenges

Several recurring features of the dictionary required dedicated handling.

3.4.1 Two-column reading order

The PDF text layer does not by itself guarantee that a full-page extraction will preserve the intended reading sequence. Reading the two columns independently prevents material from opposite sides of the page from being interleaved.

3.4.2 Abbreviated headwords inside definitions

Some definitions refer back to their own headword using an abbreviated initial form rather than repeating the full word. The parser recognizes this pattern when the initial corresponds to the record headword and reconstructs the intended internal reference.

3.4.3 Adjacent-record bleed

In a small number of cases, text belonging to the following dictionary record can remain attached to the end of the current extracted body. A cleanup rule therefore checks the tail of definitions for patterns that resemble a new entry start and removes matching residual text when the source structure indicates that a new record has begun.

3.4.4 Superscript homonym markers

The dictionary distinguishes homonymous headwords using superscript digits. These are converted into an explicit `homonym_index` field. An ASCII-digit fallback is also supported for cases where the extracted text layer represents the marker differently.

These rules are deliberately tied to the recurring conventions of the source. Exceptional or malformed entries that do not follow those conventions are preserved as residual cases rather than treated as evidence that the same parser would generalize automatically to other dictionaries.

3.5 Character Encoding and Normalization

All extracted text is stored as UTF-8. In the current version, we preserve characters as they appear in the PDF text layer and do not apply a general Unicode normalization pass.

This choice preserves the extracted form but also creates a known consistency issue. In particular, Somali apostrophes occur using more than one Unicode code point in the source, including U+0027 and U+2019. Strings that are visually and orthographically equivalent can therefore remain distinct at the code-point level.

This affects operations such as exact headword matching, duplicate detection, and cross-reference resolution. The raw representation is retained so that normalization can be introduced later without losing the source form. Unicode normalization and a normalized lookup representation are therefore planned as improvements to the resource rather than being treated as part of the present extraction.

3.6 Extraction Diagnostics and Source Comparison

We evaluate the extraction primarily through internal consistency diagnostics and direct inspection against the source dictionary.

All recognized part-of-speech codes are successfully mapped to one of the parser's internal categories, with no remaining POS normalization failures. Only **42 records (0.09%)** contain neither a definition nor a redirect.

Cross-reference resolution provides another internal consistency signal: **95.4%** of redirect records are linked to an extracted target, while **4.6%** remain unresolved.

The pair (`headword`, `homonym_index`) is not unique for all records. We observe **88 collisions (0.19%)**. These records are retained rather than automatically merged because the extraction

does not assume that repeated source records are necessarily accidental duplicates.

Across **32,801 extracted definition objects**, **96 (0.29%)** have empty text, **381 (1.16%)** contain fewer than six characters, and **48 (0.15%)** retain text matching the adjacent-entry pattern after the primary cleanup stage. We report these cases explicitly rather than silently excluding them from the resource.

In addition to these internal diagnostics, we drew a reproducible random sample of **100 extracted records** using a fixed seed and compared the parsed representation against the corresponding source material. In this sample, the reviewed fields agreed with the dictionary source.

This manual comparison should be interpreted as a source check on sampled extracted records, not as evidence that every entry in the dictionary has been recovered without error. The quality figures reported here therefore describe the observed consistency of the extraction, known residual error classes, and the results of the sampled source comparison.

4 Resource Description and Analysis

This section describes the linguistic content captured in Qaamuuska-NLP. The statistics reported here characterize the extracted dictionary records rather than general Somali language usage: the distributions reflect the editorial coverage, terminology, and conventions of *Qaamuuska Af-Soomaaliga*.

4.1 Parts of Speech and Morphological Information

The lexicon is dominated by nouns and verbs. Of the **46,314 records**, **34,726 (75.0%)** are classified as nouns and **11,445 (24.7%)** as verbs. The remaining records are distributed across a small set of other categories, including exclamations, pronouns, particles, prepositions, and numerals.

Table 2: Part-of-speech distribution.

Part of speech	Records	%
Noun	34,726	75.0
Verb	11,445	24.7
Exclamation	71	0.2
Pronoun	54	0.1
Particle	11	< 0.1
Preposition	5	< 0.1
Numeral	2	< 0.1
Total	**46,314**	**100.0**

The high proportion of nouns and verbs is also reflected in the amount of morphological information available for these categories.

Among the **34,726 noun records**, **34,570** carry a gender annotation, corresponding to approximately **99.6% of nouns**. Feminine nouns are the most frequent, with **18,410 records**, followed by **13,934 masculine nouns** and **2,226 nouns** marked as permitting either gender.

Plural information is less consistently recorded. We extract noun plural metadata for **4,698 records**, corresponding to approximately **13.5% of noun records**. This difference reflects the information explicitly encoded in the source dictionary rather than an attempt by the extraction pipeline to infer missing plural forms.

Verb morphology is highly represented in the source. Of the **11,445 verb records**, **11,444** contain a conjugation-class annotation. The four classes contain **4,170 Class I verbs**, **3,067 Class II verbs**, **2,193 Class III verbs**, and **2,014 Class IV verbs**.

The same verb records also encode transitivity. We extract **6,528 intransitive**, **4,821 transitive**, and **96 bitransitive** verbs.

These fields make the dictionary particularly useful as a source of curated grammatical information. Unlike a corpus-derived vocabulary, the records associate citation forms directly with linguistic labels assigned by the dictionary editors.

4.2 Definitions and Polysemy

A total of **25,240 records (54.5%)** contain at least one extracted definition. The remaining records are largely cross-reference entries that direct the reader to another dictionary record rather than providing an independent definition.

Across the defined records, the extraction contains **32,801 definition objects**. Most records are associated with a single numbered definition: **19,735 records (78.2% of defined records)** contain one, while **5,505 (21.8%)** contain two or more.

The maximum observed number of definitions attached to a single dictionary record is **17**, and the mean is **1.30 definitions per defined record**.

These figures provide a view of the dictionary's internal lexical organization. They should not be interpreted as an exhaustive linguistic measure of Somali polysemy, since the number and granularity of definitions depend on the editorial decisions of the source dictionary.

4.3 Homonymy

The dictionary explicitly distinguishes many same-form entries using superscript homonym markers. Qaamuuska-NLP preserves these markers as a separate `homonym_index` field.

A total of **9,212 records (19.9%)** carry an explicit homonym index.

When headwords are considered by surface string, **4,005 strings** occur in exactly two records, **411** occur in three records, and **19** occur in four or more records. The most repeated headword string appears across **16 records**.

These records are kept separate rather than collapsed into a single lexical item. This preserves distinctions made by the source dictionary and avoids treating orthographic identity as evidence that two records represent the same lexical entry.

4.4 Cross-References and Synonym Links

The dictionary contains a substantial internal reference structure.

The largest component consists of the **21,032 redirect records** described in Section 3. These records refer the reader to another dictionary entry rather than supplying an independent definition.

Of these, **20,075 (95.4%)** are successfully linked to an extracted target record.

In addition, **11,415 records** contain one or more synonym references, yielding **11,854 extracted synonym links** in total.

We keep these two relation types conceptually separate. Redirects reflect the dictionary's editorial cross-reference structure, while synonym references encode relations explicitly marked as synonyms in the source. They are therefore not treated as interchangeable semantic links.

Together, these relations provide an internal network over the dictionary records that can support lexical lookup, variant navigation, and future lexical-semantic analysis. The present resource preserves the relations as recorded in the dictionary rather than attempting to infer additional semantic relationships automatically.

4.5 Subject Domains

The source dictionary assigns subject-domain labels to a smaller subset of specialized vocabulary. We extract **1,945 domain-labelled records**, corresponding to **4.2% of the full resource**, across **16 subject domains**. Because a record may carry more than one label, these records yield **1,974 domain-label assignments**.

Table 3: Subject-domain distribution.

Domain	Records	Domain	Records
Medicine	364	Music	40
Physics	243	Politics	36
Mathematics	231	History	37
Biology	222	Commerce	14
Chemistry	221	Zoology	13
Religion	172	Law	8
Geography	155	Psychology	3
Geology	109		
Botany	106	**Assignments**	**1,974**

Scientific and technical terminology accounts for a large share of the labels. Medicine, physics, mathematics, biology, and chemistry together contribute **1,281 label assignments**. Domain counts should not be summed as unique records because multi-labelled records appear under more than one category.

This distribution reflects the source dictionary's coverage of specialized terminology and should not be interpreted as the topical distribution of Somali vocabulary as a whole.

4.6 Summary of the Resource

Qaamuuska-NLP contains several complementary layers of lexical information.

At the record level, the resource preserves the dictionary's headwords, part-of-speech distinctions, homonym organization, definitions, and cross-references. For nouns, it captures gender and a subset of explicitly recorded plural information. For verbs, it retains conjugation class and transitivity for almost the entire verbal inventory. A smaller portion of the lexicon also contains synonym and subject-domain information.

The result is therefore not simply a large word list. Its main value lies in the structured relationships between dictionary records and the linguistic information attached to them.

At the same time, these statistics describe one lexicographic source. The resource inherits the dictionary's coverage decisions, terminology, sense distinctions, and grammatical conventions. Qaamuuska-NLP should consequently be understood as a computational representation of *Qaamuuska Af-Soomaaliga*, rather than as an independent census of contemporary Somali vocabulary.

5 Potential Applications

Qaamuuska-NLP is primarily a structured lexical resource rather than a task-specific benchmark. Its fields are relevant to several areas of Somali language technology, but most of these downstream uses are not evaluated directly in this paper. We therefore describe them as potential applications and as directions for future work.

5.1 Part-of-Speech and Lexical Lookup

Every record in Qaamuuska-NLP contains a part-of-speech category derived from the source dictionary. This provides a large dictionary-based inventory of Somali words associated with grammatical categories.

Such an inventory could serve as a lexical component in POS-tagging systems, particularly for dictionary lookup, unknown-word handling, or weak supervision. It could also be used to study disagreements between lexicographic POS categories and the categories assigned to words in running text.

The current resource preserves the dictionary's own grammatical distinctions. A mapping to Universal Dependencies would make this information easier to combine with existing Somali and multilingual tagging pipelines, but that mapping is not part of the present version.

5.2 Lemmatization and Morphological Analysis

The strongest immediate use of the resource is likely to be morphological.

Noun records contain gender information and, where provided by the dictionary, plural metadata. Verb records contain conjugation class and transitivity. These fields complement existing Somali lemmatization resources, which are organized primarily around roots, derived forms, and morphological normalization [9, 10].

Qaamuuska-NLP offers a different kind of supervision: the labels come directly from an edited monolingual dictionary and remain associated with their dictionary records and definitions.

These annotations could be used as seed data for morphological analyzers, for generating training examples, or for evaluating whether models recover dictionary-defined grammatical properties. Section 6 examines one limited aspect of this possibility by testing how predictable selected labels are from headword surface form alone.

5.3 Lexical Semantics and Sense Information

The resource contains **32,801 extracted definition objects**, together with explicit homonym distinctions, synonym references, and dictionary cross-references.

These fields make it possible to study Somali lexical meaning at a level that is not available from a plain word list. Potential uses include:

- lexical similarity and synonym retrieval;
- homonym and sense-disambiguation research;
- dictionary-based semantic representations;
- terminology exploration; and
- evaluation of Somali word or sentence representations against lexicographic information.

The definitions are monolingual Somali definitions and have not been aligned with an external sense inventory or semantic ontology. They should therefore be treated as source-dictionary sense descriptions rather than as an independently validated semantic benchmark.

5.4 Terminology and Domain-Specific Vocabulary

Qaamuuska-NLP preserves subject-domain labels for **1,945 records** across 16 categories, including medicine, physics, mathematics, biology, chemistry, geography, law, and politics.

Although this is only a small portion of the full dictionary, the labels provide a useful starting point for identifying specialized Somali terminology.

Such information could support terminology extraction, educational resources, domain-specific search, or the development of specialized vocabularies for areas where Somali digital resources remain limited. Because the labels are inherited from the source dictionary, they also provide an editorially defined domain signal rather than one inferred automatically from corpus statistics.

5.5 Cross-References and Lexical Navigation

The dictionary's cross-reference structure provides another useful layer of information.

Qaamuuska-NLP contains **21,032 redirect records**, of which **20,075** are linked to an extracted target record, together with **11,854 synonym links**. These relations can support dictionary navigation and could also be used to construct graph-based views of the lexicon.

For example, a lexical search system could follow redirects automatically, expose related records, or use synonym links for query expansion. The same structure could support analyses of how the dictionary organizes spelling variants, lexical relations, and editorial references.

These relations should not all be interpreted as equivalent semantic edges. Redirects and synonym links are preserved separately because they serve different functions in the source dictionary.

5.6 Machine Translation and Language-Model Support

A structured dictionary can also provide complementary lexical information for broader language technologies such as machine translation and language models.

Definitions, morphological labels, domain information, and synonym relations may be useful for terminology handling, vocabulary normalization, lexical augmentation, or retrieval-augmented systems. In machine translation, for example, dictionary information could help with rare or specialized vocabulary even though a monolingual lexicon alone does not provide parallel sentence supervision.

Similarly, Qaamuuska-NLP could be used as supplementary lexical data when building or evaluating Somali language models, particularly for tasks involving dictionary definitions, morphology, or lexical relations.

We do not evaluate either machine translation or language-model performance in this work. These applications therefore remain possible downstream uses rather than demonstrated benefits of the present resource.

5.7 Scope

The applications above follow directly from the information represented in the lexicon, but the present paper evaluates only the resource itself and the surface-form predictability analysis in Section 6.

Qaamuuska-NLP should therefore be viewed as reusable lexical infrastructure. Its main contribution is to make the information contained in *Qaamuuska Af-Soomaaliga* available in a structured computational representation, allowing future work to test how useful that information is across specific Somali NLP tasks.

6 Surface-Form Predictability of Morphological Labels

The grammatical annotations in Qaamuuska-NLP are inherited from the source dictionary rather than generated by a predictive model. As a simple diagnostic, we ask whether selected labels show systematic relationships with the surface form of Somali headwords.

This experiment is not intended as an evaluation of a complete morphological analyzer. Instead, it measures how predictable several dictionary labels are from word form alone under a held-out split.

6.1 Experimental Setup

We consider three prediction tasks:

1. **Verb conjugation class**, using the four classes I–IV;
2. **Noun gender**, using masculine, feminine, and either-gender labels; and
3. **Coarse part of speech**, grouped as noun, verb, or other.

For each task, we use all records containing the relevant target label.

The data are divided into an **80/20 train-test split** using a fixed random seed. The partition is **headword-disjoint**: the same headword string cannot occur in both the training and test sets. This prevents the classifier from succeeding by memorizing labels associated with complete headwords that also appear during testing.

We compare two prediction strategies.

6.1.1 Majority baseline

The baseline assigns the most frequent training-set label to every test item.

6.1.2 Suffix back-off classifier

The second classifier uses word-final character sequences.

During training, it records the most frequent label associated with every suffix of length one to six characters. At test time, the classifier begins with the longest available suffix of the input headword. If that suffix was not observed during training, it progressively backs off to shorter suffixes. If no known suffix is found, it returns the majority training label.

The classifier uses only the surface form of the headword. It receives no definition, dictionary metadata, contextual sentence, external corpus, or pretrained representation.

We report accuracy and macro-averaged F1 using a fixed seed of **42**.

6.2 Results

Table 4 summarizes the results.

Table 4: Surface-form prediction results.

Task	Records	Majority accuracy	Suffix accuracy	Macro-F1
Verb class (I–IV)	11,444	0.378	**0.935**	**0.936**
Noun gender (m/f/b)	34,570	0.538	**0.864**	**0.874**
POS (noun/verb/other)	46,314	0.746	**0.869**	**0.560**

The clearest pattern appears for **verb conjugation class**. The suffix classifier reaches **0.935 accuracy**, compared with **0.378** for the majority baseline. Its macro-F1 of **0.936** also indicates that performance is not driven only by the most frequent conjugation class.

Noun gender is similarly predictable from word form. The classifier reaches **0.864 accuracy**, compared with a majority baseline of **0.538**, with a macro-F1 of **0.874**.

These results are consistent with the strong role of word-final morphology in Somali. In both tasks, relatively short suffix information contains substantial signal about grammatical categories recorded by the dictionary.

The **part-of-speech** task behaves differently. The suffix classifier improves accuracy from **0.746** to **0.869**, but its macro-F1 is only **0.560**. The difference between accuracy and macro-F1 reflects the strong class imbalance in the resource: nouns account for approximately 75% of records, verbs for nearly 25%, while the remaining categories collectively represent only a very small fraction of the lexicon.

Surface form therefore provides useful information for distinguishing the dominant noun and verb categories, but performs less well on the small residual *other* class.

6.3 Interpretation

The purpose of this experiment is modest.

It shows that several grammatical labels extracted from the dictionary are not arbitrary with respect to Somali word form. In particular, noun gender and verb conjugation class are strongly associated with character-level suffix patterns that can be learned from held-out headwords using a simple deterministic classifier.

This is useful for two reasons.

First, it provides an additional characterization of the lexical information contained in Qaamuuska-NLP. The morphological fields encode regularities that can be recovered statistically from the word forms themselves.

Second, it suggests that these annotations could provide supervision for future Somali morphological models. More sophisticated systems could combine word form with contextual, lexical, or neural representations rather than relying only on suffix lookup.

The experiment does **not** establish the accuracy of a Somali morphological analyzer on independent running text. It does not evaluate lemmatization, segmentation, contextual POS tagging, or the treatment of unseen inflected tokens in a corpus.

The results should therefore be interpreted as a **within-resource surface-form predictability analysis**, not as an extrinsic downstream evaluation.

6.4 Summary

A simple suffix-based classifier substantially outperforms the majority baseline for all three tasks, with particularly strong results for verb class and noun gender. These findings indicate that the dictionary's morphological annotations capture regular relationships between Somali word forms and grammatical categories.

At the same time, the substantially lower macro-F1 for coarse part of speech illustrates the limitations of prediction from word form alone, particularly when classes are highly imbalanced.

Together, the results provide a lightweight diagnostic of the structured labels in Qaamuuska-NLP while leaving full evaluation on independent Somali text to future work.

7 Conclusion and Limitations

We presented **Qaamuuska-NLP**, a structured reconstruction of *Qaamuuska Af-Soomaaliga* containing **46,314 dictionary records**. The resource is extracted directly from the dictionary's native PDF text layer using a deterministic, rule-based pipeline, without OCR or machine-learning-based information extraction.

The resulting records preserve a broad range of information encoded in the source, including part of speech, noun gender, verb conjugation class and transitivity, plural metadata, definitions, homonym markers, synonym references, cross-references, and subject-domain labels where these are present. The original extracted record text is retained alongside the structured representation so that parsed records can be inspected against the source material.

Rather than positioning Qaamuuska-NLP as the first or largest Somali computational lexicon, we view it as complementary to an existing ecosystem of Somali lexical and morphological resources. Earlier work has developed multi-dictionary lexical databases, finite-state morphology, corpus-linked lexicons, and purpose-built lemmatization resources. Qaamuuska-NLP contributes a different representation: a **single-source computational reconstruction of one major Somali monolingual dictionary**, with its internal lexicographic structure retained in explicit fields.

The resource also illustrates the amount of linguistic information contained in the source dictionary. Nouns and verbs account for almost the entire collection, with gender annotations available for nearly all noun records and conjugation-class information for almost all verbs. More than 25,000 records contain definitions, while redirects and synonym references form a substantial network across the dictionary.

Our surface-form prediction analysis provides a small additional characterization of this information. A simple suffix-based classifier predicts verb conjugation class and noun gender substantially better than a majority baseline, while coarse POS prediction is more affected by class imbalance. These results show that some of the dictionary's grammatical labels are strongly associated with word-form patterns. They should not, however, be interpreted as an evaluation of a complete Somali morphological analyzer or as evidence of downstream performance on independent running text.

The resource has several limitations. First, it is derived from a **single dictionary**. Its lexical coverage, sense distinctions, terminology, dialectal choices, and grammatical conventions therefore reflect the editorial decisions of *Qaamuuska Af-Soomaaliga* rather than a balanced sample of contemporary Somali usage. Qaamuuska-NLP should be understood as a computational representation of that source, not as an exhaustive inventory of the Somali language.

Second, the extraction pipeline is intentionally **source-specific**. Its rules depend on the dictionary's two-column layout, grammatical abbreviations, typography, and recurring entry patterns. The same pipeline would require adaptation before being applied to a dictionary with different conventions.

Third, the current version preserves extracted characters without a general Unicode normalization pass. In particular, more than one apostrophe code point occurs in the source. This can affect exact headword matching, duplicate detection, and cross-reference resolution. A future version could retain both the original source form and a normalized lookup representation.

A small number of extraction artifacts also remain. Of the 21,032 redirect records, **957 (4.6%)** do not currently resolve to an extracted target. The pair (`headword`, `homonym_index`) collides for **88 records**, and small fractions of extracted definitions are empty, unusually short, or retain residual text resembling the beginning of an adjacent dictionary record. These cases are retained and reported rather than silently removed.

The manual source comparison reported in Section 3 covers a reproducible sample of **100 extracted records**. Together with the internal consistency checks, this provides evidence about the observed quality of the extraction, but it should not be interpreted as a complete gold-standard evaluation of the entire dictionary.

Finally, the resource is currently **monolingual and source-specific in its annotation scheme**. The definitions are in Somali, and the dictionary's own grammatical categories have not yet been mapped to Universal Dependencies or exported through standards such as TEI Lex-0 or OntoLex-Lemon. These remain useful directions for future interoperability work.

Future work could therefore proceed in several directions: Unicode normalization and improved reference matching; closer analysis of unresolved redirects and repeated keys; mappings to established lexical standards; and downstream evaluation on independent Somali text for tasks such as lemmatization, morphological analysis, lexical retrieval, and terminology processing.

More broadly, this work shows the value of revisiting established Somali lexicographic resources as computational data. Much of the linguistic knowledge needed for low-resource NLP already exists in carefully compiled dictionaries, but often in formats designed primarily for human reading. Qaamuuska-NLP provides one example of how that knowledge can be recovered in a reproducible structured form while preserving its connection to the original source.

8 Ethics and Data Availability

8.1 Source Material and Copyright

Qaamuuska-NLP is derived from *Qaamuuska Af-Soomaaliga* [4], a copyrighted monolingual Somali dictionary published by Roma Tre Press. The source dictionary is publicly accessible as a PDF, but public access to the PDF does not by itself imply permission to redistribute its lexical content in a new structured form.

Our contribution consists of the extraction pipeline, the structured representation, and the analyses presented in this paper. Copyright in the underlying dictionary content remains subject to the rights associated with the original work.

8.2 Data Availability

At the time of writing, written permission to publicly redistribute the complete extracted lexical content is not documented in the release package. We therefore do not treat the full 46,314-record dataset as openly licensed. The complete resource and dashboard remain private research-preview artifacts while the applicable redistribution terms are clarified with the relevant rightsholder or authorized representative.

We are preparing the **extraction code**, **schema**, **aggregate statistics**, and **validation materials** for a reproducibility release through the project repository at `https://github.com/`

`Ogaal-Labs/somali_dictionary_lexicon`. The intended dashboard address is `https://qaamuuska-nlp.ogaallabs.com`; it should not be interpreted as a public full-data release until the permission status and displayed usage terms are updated.

The extraction software is intended for release under the MIT License. That software license applies only to original project software and documentation; it does not license or relicense headwords, definitions, annotations, examples, or other lexical content derived from the source dictionary.

8.3 Intended Use

Qaamuuska-NLP is intended for research and educational work in Somali language technology. The resource contains lexicographic material and does not contain personal data, human-subject annotations, or information collected from individual participants.

Acknowledgments

The Qaamuuska-NLP reproducibility materials are being prepared in the private GitHub repository. The intended address for the accompanying web dashboard is qaamuuska-nlp.ogaallabs.com. We acknowledge Ogaal Labs for supporting the development and presentation of these project resources.

References

[1] Abdillahi Nimaan. Building and evaluating somali language corpora. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 73–76, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[2] Abdisalam Badel, Ting Zhong, Wenxin Tai, and Fan Zhou. Somali information retrieval corpus: Bridging the gap between query translation and dedicated language resources. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7463–7469, Singapore, December 2023. Association for Computational Linguistics.

[3] Silvia Piccini, Giuliana Elizabeth Vilela Ruiz, Andrea Bellandi, and Enrico Carniani. Tracing linguistic heritage: Constructing a somali-italian terminological resource through explorers' notebooks and contemporary corpus analysis. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-Resourced Languages at LREC-COLING 2024*, pages 357–362, Torino, Italy, May 2024. ELRA and ICCL.

[4] Annarita Puglielli and Cabdalla Cumar Mansuur, editors. *Qaamuuska Af-Soomaaliga*. Roma TrE-Press, Rome, Italy, June 2012.

[5] Jama Musse Jama. Somali corpus: State of the art, and tools for linguistic analysis. Manuscript, Redsea Cultural Foundation, 2016.

[6] Annarita Puglielli. The somali language lexical project: Aims and methods. In Martin Pütz, editor, *Discrimination through Language in Africa? Perspectives on the Namibian Experience*, pages 123–137. Walter de Gruyter, Berlin and New York, 1995.

[7] Nick Thieberger and Nadia Faragaab. Developing a somali dictionary application. *Bildhaan: An International Journal of Somali Studies*, 14(1), 2015. Article 10.

[8] Abdullahi Mohamed Jibril and Abdisalam Mahamed Badel. Creating and analyzing a dictionary-based lexical resource for somali homograph disambiguation. In Arun Sharma and Ritu Rani, editors, *Artificial Intelligence and Speech Technology*, volume 2390 of *Communications in Computer and Information Science*, pages 265–276. Springer, Cham, 2025.

[9] Shafie Abdi Mohamed and Muhidin Abdullahi Mohamed. Lexicon and rule-based word lemmatization approach for somali language. In *AfricaNLP Workshop at ICLR 2023*, 2023.

[10] Abdifatah Ahmed Gedi, Shafie Abdi Mohamed, Yusuf A. Yusuf, Muhidin A. Mohamed, Fuad Mire Hassan, and Houssein A. Assowe. Morphologically-informed somali lemmatization corpus built with a web-based crowdsourcing platform. In *Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026)*, pages 179–189, Rabat, Morocco, March 2026. Association for Computational Linguistics.

[11] Jennifer Tracey, Stephanie Strassel, David Graff, Jonathan Wright, Song Chen, Neville Ryant, Seth Kulick, Kira Griffitt, Dana Delgado, and Michael Arrigo. Lorelei somali representative language pack. Linguistic Data Consortium, LDC2026T03, 2026.

[12] HaBiT Project. Somali web corpus (sowac16). Deliverable D3.4a, 2016.

[13] Burcu Karagol-Ayan, David Doermann, and Bonnie J. Dorr. Acquisition of bilingual mt lexicons from ocred dictionaries. In *Proceedings of Machine Translation Summit IX: Papers*, New Orleans, USA, September 2003.

[14] Elvis Mboning, Daniel Baleba, Jean Marc Bassahak, Ornella Wandji, and Jules Assoumou. Ntealan dictionaries platforms: An example of a collaboration-based model. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 66–72, Marseille, France, May 2020. European Language Resources Association.

[15] Piotr Bański and Beata Wójtowicz. A repository of free lexical resources for african languages: The project and the method. In *Proceedings of the First Workshop on Language Technologies for African Languages*, pages 89–95, Athens, Greece, March 2009. Association for Computational Linguistics.

[16] Michael Maxwell and Aric Bills. Endangered data for endangered languages: Digitizing print dictionaries. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 85–91, Honolulu, Hawaii, USA, March 2017. Association for Computational Linguistics.

[17] Mohamed Khemakhem, Luca Foppiano, and Laurent Romary. Automatic extraction of tei structures in digitized lexical resources using conditional random fields. In *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017*, pages 598–613, Brno, Czech Republic, 2017. Lexical Computing.

[18] Jack Bowers, Mohamed Khemakhem, and Laurent Romary. Tei encoding of a classical mixtec dictionary using grobid-dictionaries. In *Electronic Lexicography in the 21st Century: Proceedings of eLex 2019*, pages 397–416, Sintra, Portugal, 2019.

[19] Murat Jumashev, Aida Kasieva, Gulnura Dzhumalieva, Aigul Tursunova, Meerim Ryspakova, and Jonathan North Washington. Structured data from dictionary text: Applying llms for low-resource cross-lingual information extraction. In *Analysis of Images, Social Networks and Texts*, volume 2364 of *Communications in Computer and Information Science*, pages 18–32. Springer, 2025.

[20] David Setiawan, Temuulen Khishigsuren, Milind Agarwal, Pagnarith Pit, Aso Mahmudi, and Ekaterina Vylomova. Mudidi: A two-stage framework for multilingual dictionary digitization with language models, June 2026.

[21] Johanna Cordova and Damien Nouvel. Toward creation of ancash lexical resources from ocr. In *Proceedings of the First Workshop*

on Natural Language Processing for Indigenous Languages of the Americas, pages 163–167, Online, June 2021. Association for Computational Linguistics.

[22] Ben Bongalon, Joel Ilao, Ethel Ong, Rochelle Irene Lucas, and Melvin Jabar. Using open-source tools to digitise lexical resources for low-resource languages. In *Electronic Lexicography in the 21st Century: Proceedings of eLex 2021*, pages 92–106, 2021.

[23] Piotr Bański, Jack Bowers, and Tomaž Erjavec. Tei-lex0 guidelines for the encoding of dictionary information on written and spoken forms. In *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017*, pages 485–494, Brno, Czech Republic, 2017. Lexical Computing.

[24] John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. The ontolex-lemon model: Development and applications. In *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017*, pages 587–597, Brno, Czech Republic, 2017. Lexical Computing.

[25] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia, May 2016. European Language Resources Association.