The Need for a Novel Approach to Design Derivation Lexicon for Semitic Languages

Enchalew Y. Ayalew[1], Laure Vieu[2] and Million M. Beyene[3]

[1]Software Engineering Department, Addis Ababa Science & Technology University, Ethiopia
enchalew.yifru@aastu.edu.et
[2] National Center for Scientific Research (CNRS), Institut de Recherche en Informatique de Toulouse, France
Laure.vieu@irit.fr
[3] School of Information Science, Addis Ababa University, Ethiopia
million.meshesha@aau.edu.et

Abstract. Morphology knowledge is relevant in language learning, information retrieval and natural language processing. Derivation lexicons are organized and comprehensive collections of the morphological variants of a language's vocabulary. These lexicons can be developed either through analysis-based synthesis of large text corpora or through synthesis of surface forms from roots, stems, lemmas and morphological rules. Much of the research attempted in developing derivation lexicon for Indo-European languages, which are concatenative, focus on analysis-based synthesis, as they do have well-developed preprocessing tools and organized text corpora. However, the methods for these languages are not appropriate for non-concatenative languages such as Semitic languages. Moreover, most of the Semitic languages, except Arabic and Hebrew, do not have well-developed text corpora and language processing tools. Hence, a novel approach that can cater for the root-pattern and rich morphology of these languages is necessary. This paper is therefore a comprehensive survey of the literature, an analysis motivating an innovative and generic morphological synthesis approach with illustrated architecture. It is part of a larger project tailored for designing an innovative, generic, approach to derivation lexicon development for Semitic languages.

Keywords: Semitic Computational Morphology, Lexicon Design, Derivation Lexicon

1 Introduction

Lexical ontologies such as dictionaries, thesauri or WordNets have been used to improve various computational applications, such as information retrieval (IR) and natural language processing (NLP)[21]. Yet, these resources remain short of addressing certain NLP and IR tasks. This motivated researchers to build derivation *lexicons*, i.e., organized clusters of part-of-speech (POS) variants; the "categorical variation of a word with a certain part-of-speech is a derivationally-related word with possibly a different part-of-speech" [17, p.1]. In an online version of CatVar 2.0, a derivation lexicon for English, the POS variants of 'break' are categorized into 39 clusters[1], indicating the word, the POS and the source lexicon in tabular form. For example, two clusters of "break" include $break_{N}$, $break_{V}$, $broken_{Adj}$, $breaker_{N}$, $breakage_{N}$, $breakers_{N}$, $breaking_{N}$, and $breaking_{Adj}$ as one cluster; and the second which encompasses $breakable_{N}$, $breakable_{Aj}$, $breakability_{N}$, $breakableness_{N}$. [17] Noted that link-ability[2] principle was used to create clusters (see details on link-ability in section 4.2).

As we shall see below, due to the successful use of morphological information in improving NLP and IR tasks, many language-specific derivation lexicons have been developed and are under development. Yet, to our knowledge, there are no derivation lexicons for any of the Semitic Languages; and hence in this paper we propose a novel approach suited to design derivation lexicon for Semitic Languages.

The rest of the paper is organized as follows. In section 2, we briefly introduce Semitic Languages and elaborate on their shared features. Section 3 looks at the major applications of derivation lexicons. Section 4 describes common approaches to non-concatenative, including Semitic languages, morphological computations and existing approaches to derivation lexicon development. Section 5 presents our proposed approach and finally, in Section 6 future work directions are highlighted.

2 Overview of Semitic Languages

Semitic languages belong to the Afro-Asiatic family of languages which mainly cover areas such as the Middle East, North and East Africa. They used to be spoken, as back in time as the 2500BC, by populations who lived in areas spanning from Ethiopia, Sudan and Saudi Arabia in the south to today's Syria in the North and what is known today as Iran and Iraq in the East [39].

According to [39], these languages are grouped into East and West Semitic. While the East Semitic languages spoken today are Amorite and Eblaite, the West Semitic group has more diverse languages and hence is sub-classified as Central and Southern Semitic. Surviving languages in the Central Semitic sub-group are Arabic and North Western Languages such as Hebrew and Aramaic. The languages that constitute the Southern Semitic sub-group are South Arabian (spoken in Yemen and Oman) and Western South Semitic which are spoken in Ethiopia. The Ethio-Semitic cluster con-

[1] https://clipdemos.umiacs.umd.edu/catvar/, last accessed 2021/12/03.
[2] Link-ability[17] is the percentage of word-to-word links resulting from a specific source.

stitutes the Northern (such as Tigre, Tigrinya and Geeze-an old liturgical language) and Southern (Amharic, Harari, Gurage, Chaha and Gura) languages. [5] Also included Argoba to the list.

The most widely-spoken Semitic languages today are Arabic, Amharic, Hebrew, Tigrigna, Syriac and Maltese-an Arabic dialect influenced by Italian language [45], Modern Aramaic, Mandaic and different dialect of Modern South Arabian languages [39]. [45] Noted that Arabic and Amharic are-respectively-the first and second widely spoken Semitic languages. Hebrew is the fourth widely spoken language, next to Tigrigna, and is also relatively well researched like Arabic.

2.1 Common Features

Languages that have closely-related features enable not only in speeding up language learning but also in sharing and adapting computational solutions easily. This is true of Semitic languages as they are related in their phonology, morphology, lexicon and syntax [39]. Narrowing our focus on morphology and lexicon tells us that these languages have complex morphologies and rich verbal lexicon. In addition, words of the same root have related semantics. These common features are illustrated taking Arabic, Amharic and Hebrew as examples.

Complex Morphology. Word formation in Semitic languages is so complex that it intensively involves both inflection and derivation. While derivation is mainly non-concatenative, inflection is dominated with suffixation and prefixation in addition to the reduplication of certain characters in the stem [4]. Inflection is meant to create variants of a lemma in the same syntactic category (showing person, number, tense, gender, etc). On the other hand, derivation helps to create new lemmas with different syntactic categories (nouns, verbs, adjectives or adverbs) from verbal roots, patterns and linguistic rules. While roots are generally consonant characters, the pattern is a sequence of consonant-vowel (like CVC, CVCC, CVCVCV …) forms where actual consonant and vowel characters will be inter-digitized using rules to form stems or lemmas. When stems/lemmas are inflected, they form surface forms.

For example, in Amharic from the tri-literal root '*m-k-r*': standing for 'advise', we can generate plenty of POS variants. To mention some, CVCC: "mIkr$_N$"[3]-advice, CVCVCV: "mekari$_N$"-adviser, CVCCVCV: "mekkere$_V$"- advised, C-CVCCVCV: "te-mekkere$_{Adj}$"- advised or C-CVC[C] VCV: "te-mek[k]ari$_N$"-advisee, V-CCVCVC-V: "a-mmakari$_{Adj}$"-advisory, V-CCVCVC-V: "a-mmakari$_N$"-advisory, etc.

Lexicon Rich in Verbs. The complexity of word formation in Semitic languages' leads to a vocabulary rich in verbs. For instance, about 75% to 80 % of Amharic dictionary entries consist of verbs or their de-verbal nouns/adjectives and hence an "exhaustive verb list is a substitute for a complete dictionary… the verb is the language" [43, p.73]. There are even a larger proportion of verbs in the Arabic vocabulary than

[3] Romanization is based on: The System for Ethiopic Representation in ASCII by Yitna Firdyiwok and Daniel Yacob(1997).

in Amharic, i.e., "verbal roots and their derivative nouns and participles make up 80% to 85% of all Arabic words"[4]. It is also noted that "Hebrew is primarily a verbal language" and "every Hebrew verb (and every noun) is based on a three-consonant root … which encodes the basic semantic meaning or purpose of a given verb or noun"[5].

All verbs and most nouns have roots [39]; hence a verbal root serves to produce not only verbs but also nouns. In addition, adjectives are also produced from verbal roots [4], in a manner similar to nouns [13]. However, prepositions, conjunctions, simple nouns and adjectives, also known as "primitives" [9], are not derivable from verbs. For example, in Amharic, the noun '*bEt*: *house*' and the adjective '*blIh: smart*'are not derivable from verbs [4]. In general, a Semitic lexicon built based on verbal roots covers most of a language's vocabulary entries.

Semantic Relatedness. Words derived from a given root are, broadly speaking, related in meaning [9, 39]. [9] Indicated that the Arabic words *kataba*, *kaataba*, *maktabun*, *maktabatun*, *kitaabun*, *maktuubun*, and *kuttaabun* are derived from the same root *k-t-b*, representing not only similar morphological and phonological relation but also, at various degrees, similar semantic contents such as the semantic meaning of *writing*. The equivalent English meanings respectively are *write*, *correspond*, *office*, *library*, *book*, *destiny*, and *Koran school*, which are not morphologically related. Except *destiny*, the rest of the terms are strongly related in meaning. Destiny is unrelated in meaning from the rest because it might be the result of "semantic drift" –deviation mainly due to "usage over time" [11]. The Hebrew root '*k-t-b*' produces surface forms which are related in meaning like '*ktb'/katav/:* write, '*hktib'/hiktiv*/: dictate,'*hktbh'/haktava*/: dictation, *mktbh*: writing desk[17]. Similarly, the Hebrew words *zimər* ('sing'), *zamar* ('singer') and *zimə*ra ('singing) are derived from the root *z-m-r* [34].

In Amharic *sella* ('sharp, have keen edge'), *sale*('sharpen'), *selessele* ('wear thin, weak'), *sellele* ('become paralyzed, withered') with a common meaning of slender are derived from same root '*s*-l'[8][6]. The Arabic root '*k-t-b'* does have the equivalent Amharic root *S-h-f* [5, p.122]. It is derived from the Geez *Sehafe*[7], which in Amharic means *Safe*, referring to "he wrote". Derivations from this root include *meShaf* ('Book'), *meSaSaf* ('correspond'), *meSaf*('write'), *Shuf/Sfet*/ (anything written or inscription), *Sehafi* (writer), *meSafia*('instrument for writing'), *aSaSaf* ('manner/style of writing'), *maSaf* ('dictate').

Although words of the same root do have one shared lexical meaning [9] and the root stands out as the core lexical content of words [34], the root represents one aspect of lexical meaning shared by the derived stems [8]; the rest of a word's meaning for these languages come from templates [9].

---

[4] https://www.memrise.com/course/110178/1500-arabic-verbs-by-frequency/, last accessed on 2021/03/01

[5] https://www.hebrew4christians.com/Grammar/Unit_Ten/Introduction/introduction.html, last accessed 2021/03/04

[6] '*s-l*' is not in a root corpus of [5, p.111]; instead there is '*sll*': gloss 'be paralyzed'.

[7] Amharic-English Dictionary by Kane (1990), Vol.1, pp.2249.

3 Application of Derivation Lexicons

Language learning and computation (IR and NLP) equally benefit from using morphology information. The first sub-section illustrates benefits for language learning; the second and third sub-sections address the merits of morphology in computation.

3.1 Language Learning

Derivation morphology knowledge speeds up the ability of children to learn new vocabulary [7] and empower the analysis and understanding of language learning from infancy to adulthood [26]. It enhances second language learning [12], helps in reading and spelling accuracy [1], a key to the access and construction of sentence syntactic structure as well as organizes internal lexicons [41].

Language learners detect and understand morphological variants easily by analyzing words into their morphemes/morphological sub-structures/, particularly detecting the root in the derived forms and then give definitions on the basis of the root [7].

3.2 Information Retrieval (IR)

Effective IR systems allow the retrieval of documents in a collection that match user queries. However, IR systems are unable to fully address user information need either due to polysemy—a situation where there are multiple possible meanings for a word/phrase (e.g., bank-'financial organization' vs. bank-'river side') —or synonyms—multiple words having the same meaning (e.g., student vs pupil). While polysemy may confuse the IR system to retrieve irrelevant documents, synonymy may not allow the retrieval of all relevant documents in a collection.

Morphological variants are relevant both at indexing and querying time to address part of these problems. During document indexing morphological variants are conflated to a single indexing term, thereby allowing the retrieval of all possible documents with the variants. During querying time, users can reformulate their search by considering system suggested morphological variant alternatives.

In general, morphology information enhances IR through query expansion and conflation-based document indexing [31]; CatVar is relevant for IR research [19].

3.3 Natural Language Processing (NLP)

In NLP, morphology knowledge is useful in machine translation, spell check, lexicon compilation, POS tagging and sentence construction [23]. CatVar improves natural language generation and machine translation [17], textual and lexical entailment [6], helps to enhance and induce semantic role resources for predicates of nouns [29] and paraphrase identification [30]. It has also the potential in lexicon construction and enhancing WordNets [17]. The German DEriveBase is used in improving similarity prediction and synonym choice [33].

4 Approaches to Derivation Lexicon Development

Morphological computations are tailored either to form a word from the parts, i.e., synthesis/generation or to break up a word into its components, i.e. analysis. Most NLP and IR research focus on analysis, by taking words of a text and breaking them up into their components and whenever necessary reproducing words from these components, thereby merging analysis and synthesis together [13]. When the goal is to develop a lexicon, however, analysis is conducted on finite text corpora and thus unable to produce comprehensive vocabulary entries. This is, particularly, true for most Semitic languages which do not have large-sized, organized corpora. Moreover, analysis involves using tools for a number of pre-processing phases such as sentence detection, tokenization, POS tagging and stemming or lemmatization.

On the other hand, synthesis takes on a collection of finite 'primitive' linguistic components—i.e., roots, stems or lemmas, along with linguistic rules— and then build words resulting in a comprehensive coverage of a language's vocabulary. This is more so for Semitic languages, where word formation is based on the interdigitations of consonantal roots with vowels, based on patterns, as explained in section 2.

Both synthesis and analysis-based synthesis have the downside of producing out-of-vocabulary words. Machine learning—decision trees implemented in Weka[8]—technique was used to reduce invalid word entries as was the case in [35]. We plan to test a similar method. Alternatively, we also planned to experiment on a less data intensive valid word prediction method. This method should depend only on small learning seed data—instead of large corpora— from which the required "full valid" vocabulary of a language is built. This is the direction adopted in this project.

In the remaining subsections, we first briefly look into non-concatenative finite state morphology (FSM). It is considered appropriate for Semitic Language morphology processing. This is followed by the specific approaches used in developing derivation lexicons for Indo-European languages; they focus mainly on analysis-based generation. Lastly, we look at the approaches used in generating various lexical resources for Semitic languages.

4.1 Non-concatenative Finite-State Morphology

The concept of finite-state morphology was proposed in the early 1980s. It was conceived to overcome the computational difficulties of morphologically complex languages in general. The idea was first tested on the Finish Language [25] following a "Two-Level Morphology" approach. The model is based on a lexicon, set of two-level rules processed in parallel and a small set of finite state automata (FSA) [39]; it handles both analysis and generation. However, this approach was not sufficient for Semitic languages which require more than two levels of representation.

Hence, [22] proposed multi-level implementation based on the theory of auto-segmental [28] approach to Semitic morphology processing. It outlined a quadruple-

[8] https://www.cs.waikato.ac.nz/ml/weka/, last accessed 2021/03/01

tape finite state machine to describe the independent morphemes of Arabic, making it more palatable for other Semitic languages as well. However, the rules to control tape manipulation were arbitrary. Therefore [24] came up with "attractive rule" control of the four-level tapes by applying it to Arabic and Syriac languages.

Later on, [15] extended [24]'s idea by adding a fifth tape, with the goal of developing the Arabic morphological analyzer and generator named MAGEAD. The five levels have different purposes [15]: level l represents patterns and affixation morphemes, level 2 stands for roots, level 3 stands for vocalism, levels 4 and 5, respectively, stand for phonology and orthography. However, it is learnt that the complexity of transitions between levels exponentially increases with the number of tapes [20]. Therefore, [19] tried to simulate the representation of multiple levels with a single-tape, claimed to be realizable on available standard finite-automaton toolkits. On the other hand, the single FSA simulation has resulted in search speed limitations during surface forms generation and search-based analysis [20]. This indicates that simulating multiple tapes into a single tape simply makes the problem cyclic.

To tackle the inherent efficiency-problem of FSA for multi-tape representation, a Finite-State Registered Automata (FSRA) was proposed [10]. It involves supplementing existing FSA with finite memory/registers so as to save space. The registers are made small in number and help to avoid the need to repeat paths in order to memorize a finite set of symbols. FSRA is efficient, reducing the quadratic time O(r*p) for traversing arcs in an ordinary FSA to linear time O(r+p), where r and p are the number of roots and patterns, respectively.

4.2 Indo-European Languages

Indo-European languages have concatenative word formation morphology. These languages rely on analysis based synthesis as the dominant approach to derivation lexicon development. This may be attributed to the availability of sufficient corpora, effective preprocessing tools and the relative "simplicity" of reducing words to basic forms—stems or lemmas. In the following paragraphs, we describe the most influential derivation lexicon development research for this group of languages.

A suffixation-based probabilistic unsupervised machine learning technique was used to strip off suffixes from words of an inflectional lexicon aiming to produce French's derivational families [14]. The main intuition to clustering words to relational families is to add words into a family as long as they are 'p' similar and relate them with suffix pairs. This is assuming suffix pairs from different families don't co-occur. 'P' stands for the number of similar sequenced characters between derived forms. The intuition is implemented using hierarchical agglomerative clustering and minimum/maximum spanning tree graphs respectively, for identifying derivational families and suffixation. There is no evidence if effort was made for sub-clustering of a word's variants; the proportion of any singleton clusters in the result is not reported.

[17] Developed a large-scale categorical database for English, known as CatVar. It was based on pre-existing data sources and tools: corpora, tree banks, lexicons and stemmer. The clustering of derived and inflected forms is based on three link-ability concepts: natural link-ability (pairs of words whose form doesn't change across cate-

gories like zip_V, zip_N), Porter link-ability (words linkable by reduction to a common Porter stem) and CatVar link-ability (link-ability of two words appearing in the same CatVar cluster). It is reported that near to half of CatVar's clusters are singleton entries. Of which, 75% are nouns and one-fifth is adjectives. Unlike this, derivationally related senses/forms in the manually built wordNet[9] consist of two or more POS variants (noun-verb, noun-adjective, verb-adjective; noun-verb-adjective). Unless a cluster has at least two POS variants, the lexicon becomes a simple word list with little purpose.

Inspired by the applicability of CatVar in IR and NLP tasks, [44] developed a lemma-based derivational knowledge base, i.e., DERIVBASE, for German using a large German web corpus, pre-existing POS tagger, parser and lemmatizer. The induction of derivational families for nouns, verbs and adjectives was based on rules from text books. The rules capture intra-POS and inter-POS derivations from POS-tagged lemmas and paradigms (zero-derivation[10], prefixes, suffixes, circumfixes and stem changes). Derivation rules are set for POS-pairs as: N-N, N-A, N-V, A-A, A-V, V-V. The clustering rule is formulated in such a way that a binary derivation relation between two lemma-paradigm pairs is considered valid if the second pair can be derived from the first one. Of the total 239680 derivational families, 17799(around 7.4%) reported to be non-singleton clusters where as the majority (most reported to be compound nouns), i.e., 221881 (>92%) are singleton clusters. [40] And [42] were motivated by the outcome of DERIVBASE and hence used similar methods in developing the derivation lexicons for Croatian and Russian Languages respectively.

In general, singleton clusters are the major drawback of both CatVar and DERIVBASE. This important problem calls for incorporating innovative intuitions rather than relying only on the link-ability concepts of CatVar or a single intuition as is the case in DERIVBASE or French's derivational families. Considering multiple—possibly hierarchical— intuitions including linguistic once can reduce the problem.

4.3 Semitic Languages

Our effort to review the literature on derivation lexicons for Semitic Languages reveals that such resources are not yet in place. However, we found several purpose specific lexical resources; showing derivation as an important approach for resource building. In this sub-section we look at these efforts. The generation-based methods (including the analysis-based generation) help us to learn about achievements and gaps on derivation lexicon development. Our discussion excludes any manually-developed resources including the lexicons of BAMA[11] and SAMA[12].

[3] Developed a large Arabic lexicon for use in an open source FST-based, bidirectional analyzer and generator for Arabic known as AraComLex 2.1[13], with the goal of overcoming the obsolete entries in SAMA lexicon. It was based on Arabic Gigaword

[9] https://wordnet.princeton.edu/download/current-version, last accessed 2021/06/24
[10] Zero-derivation results in POS variants of identical forms (e.g, the farm => to farm;)
[11] Buckwalter Arabic Morphological Analyzer
[12] Standard Arabic Morphological Analyzer
[13] AraComLex 2.1. is an open source, lemma-based, analysis and generation FST for MSA.

corpus and news articles from the Al-Jazeera web site. MADA[14] was used for pre-annotation such as to lemmatize, diacritize, POS-tag and disambiguate the input data. The use of a multilayer perceptron machine learning in Weka enabled AraComLex 2.1 to have better coverage from its predecessor, i.e., AraComLex 1.0, but slightly lower in coverage from that of SAMA.

[2] Proposed a simple, rule-based, algorithm with the goal of developing an inflectional analysis-based Arabic word generator based on input word from the user. The input is analyzed to its stem and inflectional components. These components then go through generation.

Unlike [2][3], the researches discussed hence forth are based on synthesizing morphological components as input. The early attempt along this line is [43] which is a rule-based approach used to generate surface forms alternatively either from Amharic roots or perfect & infinitive forms. The goal was to determine which of the three predicts the two others best. This was used to understand the impact of such derivation in language acquisition/learning. Morphology rules, applied to the 42 Amharic verb-classes of [5], along with the 1280 Amharic roots from the same source were implemented in BASIC. It is shown that the program with the root as input was capable of correctly deriving infinitives of all verbs using eight rules; and through the infinitive, all other verbal forms, indicating that the root has the advantage as it predicts the other verbal forms unambiguously.

[23] Developed a synthesizer for Amharic verb forms from manually compiled 145 tri-literal perfective verb roots. Derivation rules, suffixation, vowel and consonant change as well as platalization rules were considered as input to the synthesizer. The synthesizer was complemented with a neural network method to predict new roots not available in the root database. The limitations of this study are too shallow testing and focus only on perfective verbs, among others. Hence, the report included several recommendations like the need to consider other verb types and develop full-fledged synthesizer that can be applied in machine translation, spell checking, lexicon compilation, POS tagging and sentence construction, among others.

[36] Developed an XFST-based morpho-graphemic rule-based model generator/analyzer for Amharic nouns (loan and native nouns included), verbs and adjectives. The generation component accepts roots (from bi-radical to quad radicals) as inputs undergoing respectively through vowel intercalation, concatenating input affixes, handle phonological alterations and finally generate the grapheme form. Similarly, [38] also developed an XFST-based analyzer/generator to produce Amharic verb-lexicon for use in a machine translation experiment. Report shows the generation of main verbal-forms from three-to-five radical roots using rules. The report doesn't show the derivation of nouns and adjectives. Both [36] and [38] indicate neither the size and source of roots and rules nor the volume of records created.

[16] Developed a large-scale lexeme-based Arabic morphological generation system known as Aragen. It is based on the database (prefix, stem, and suffix) of BAMA with a new engine performing generation instead of analysis. Input data was extracted from the UN Arabic-English corpus. The baseline generator used a simple concatena-

[14] MADA- stands for Morphological Analysis and Disambiguation for Dialectical Arabic

tive word structure rule and a small lexicon with 70 entries. Evaluation report shows that Aragen's performance was much better than the base line both in under-generation and over-generation errors.

[37] Developed a rule-based morphological generator for Arabic with the goal of using it in an interlingua-based machine translation project for spoken dialogues. It generates inflected nouns/adjectives, verbs and particles from stems and grammatical features. Although it is indicated that evaluations were satisfactory, it was not quantified to understand the extent of the system's proformance. The generator is reported to have been successfully used in other applications such as Arabic-Audio Indexing and intelligent computer-assisted learning for Arabic.

Intending to semi-automatically extend AWN (Arabic WordNet), [35] used lexical rules, as regular expressions, to produce derived forms (nouns, verbs and adjectives) from the roots of 2296 verbs in Arabic WordNet. To filter out over-generated forms, decision tree classifiers in Weka toolbox were used. The learning was based on Arabic Gigaword corpus (to get relative frequency of each inflected form), Arabic NMSU dictionary entries (to check presence of base form and its POS tag) and positive /negative examples.

[13] Developed a tool named HornMorph for the analysis and generation of two Semitic languages (Amharic & Tigrigna) and a Cushitic language (Oromo). Focusing on the generation component of Amharic and Tigrigna, rules as FSTs were implemented in python. Input included 1851 verb roots from dictionary and 6471 noun stems (for Amharic) and 602 verb roots for Tigrigna.

[32] Has synthesized a lexicon of 15,400 Arabic verbs into 2.5 million inflected forms. Input verb lemmas, grouped into 31 root classes, were extracted from a pre-compiled full-form (with diacritic markers) dictionary. For each lemma class, inflection rules were implemented using FSTs to produce the surface forms.

In a PhD research, [27] used a rule-based model, to synthesize Arabic verbal lemmas and inflected forms from roots and templates in an effort to develop a morphological analyzer and generator. An input lexicon of 15452 verb lemmas-for 3706 roots-was used to generate a lexicon of more than 1.68 million verbal inflected forms.

[18] Derived the surface forms of twenty simple present tense Amharic verbs that begin only with consonants. It is based on the theory of network morphology, implemented using the partially object-oriented tool, DATR. The twenty verbal templates followed four stem patterns such as CVCCVCV, CVCVCV, CV and CVCVCVCV. The system handled the addition and deletion of phonological changes, which happen when the sixth order 'I' in a stem follows the consonants 'd, n, r, z or l'(alveolars), deletion occurs. On the other hand, when the fourth order form follows m, b, l, r, g, q, t or c, it changes to a third order form and 'a' is added (addition takes place). Both changes also cause change in the root and pattern template.

The assessments of the aforementioned papers on Semitic languages show that derivation of inflected verbal, adjectival or noun forms are possible from verb roots using linguistic rules. [43]'s use of the Amharic derivation rule set and pre-existing root corpora from [5] to derive various verbal forms; and [13]'s effort in developing an FST for both analysis and generation of Amharic and [Tigrigna] is an encouraging input for Amharic lexicon development research. [23]'s attempt to synthesis Amharic

verbal forms from roots, though limited, and the outlook towards suggesting for a broader research to develop full-fledged synthesizer for use, for instance, in Amharic lexicon compilation, among others, is an important point. Furthermore, [36]'s use of FSs to generate Amharic POS variants (nouns, verbs and adjectives) and [38]'s derivation of verbs from roots and rules further justifies the viability of root-based derivation in Amharic and an important input to resource development.

In general, it is important to note that the efforts on Amharic [Ethio-Semitic in general] morphological analysis and generation— including those not reviewed in here due to space economy— are at an experimental stage at large. They are not to the level of building a lexicon of derived forms accessible even for research let alone for general public use.

From among the papers on Arabic language synthesis-based derivation, the ones by [27] and [35] have important ideas to consider particularly in their use of organized rule-set and roots to generate verbs, nouns and adjectives. Moreover, [35]'s noise filtering approach is also reported to be effective and hence is encouraging. However, most Arabic generation research reviewed such as [2, 3, 16, 32, 37] focused on inflection than generation using existing tools and resources.

Other than generating surface forms for specific uses or as a research in its own right, none of the attempts discussed on Semitic languages' utilized rule and root-based generation to produce organized lexical resource similar to CatVar or DERIVBASE. Thus, no plan or effort is reported in linking (clustering) POS variants (nouns, adjectives, verbs) of a root into meaningful categories. The main component of research in derivation lexicon development is the clustering of POS variants using various methods like machine learning [14], link-ability [17] and rule-based [44].

5 Proposed Approach

Our understanding—thus far—indicates three important points. Firstly, the morphology of Semitic languages is quite different from Indo-European languages. While the derivation and inflection morphology in the later is concatenative, the former has predominantly non-concatenative derivational and inflectional morphology. This is a challenge to utilize or adapt available NLP tools and algorithms for Indo-European languages to Semitic languages. Secondly, most Semitic languages suffer from the lack of accessible and well-functioning resources and tools for language processing. Thirdly, most of the research in language processing focuses on morphological analysis or analysis-based synthesis. However, we recall that there are few efforts, particularly for Semitic languages, which focused on synthesis-based resource development. This warrants that morphological synthesis— as an important approach in resource development— is of an interest on its own. Moreover, synthesis involves less preprocessing language tools making it more appropriate for most Semitic languages.

Semitic languages' research in the generation of lexical resources hasn't yet given any attention to derivation lexicon development. Our focus is, then, to advance the existing Semitic derivation morphology research a step further. It is to design a generic approach that synthesizes words from roots using rules, cluster POS variants and

POS-homogeneous forms of a root; note that POS homogenous clustering is the main organizing principle of WordNet's entries.

Generally our approach is novel in that it is the reverse of the lexicon development approaches used in resource-rich languages. For instance, in CatVar or DERIVBASE, the idea is analysis-based generation; mainly relying on pre-existing resources and tools. This limits the accuracy of results due to cascading of preexisting resources limitations into the generated lexicons. Moreover, the use of limited clustering intuitions has resulted in lexicons dominated with singleton members. To overcome this, the intent is to use multiple intuitions and language features.

For instance, POS variant clustering can be achieved using multiple, hierarchical heuristic insights. For instance, integrating root signature for each derived-form produces at-least one macro-cluster with multiple elements. Given, a macro-cluster, integrating alternative, possibly multiple intuitions can result in more coherent sub-clusters. One of the intuitions can be setting the threshold for sub-cluster members to be a minimum of two; otherwise, the candidates for sub-clustering remain to be members of the macro-cluster.

Unlike POS-variants, POS-homogenous clustering is mostly achievable with linguistic rules and hence is anticipated to be handled as such. Features for both POS-variant and POS-homogenous clustering can be captured at the time of derivation.

Our approach has five steps, as represented in (Fig.1).

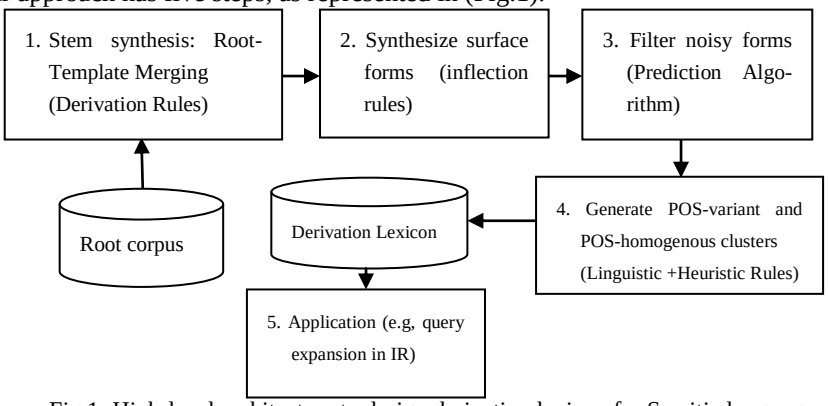

Fig.1: High-level architecture to design derivation lexicon for Semitic languages

We now illustrate our design highlighting on important points. Implementation of the approach is based on Amharic, the second widely spoken Semitic language. Our intent is rule-based stem and surface form synthesis. Rules from text books (Baye, 2009; Bender and Fulas, 1978) and the root corpus (around 1280) of Bender and Fulas(1978) are considered.

Our design excludes simple nouns (e.g., semay: 'sky'), adjectives (qey: 'red'') and adverbs. Instead, it focuses on verbs, de-verbal nouns and adjectives which take up the lion share of these languages' vocabularies.

Rule implementation is based on regular expressions or finite states in general. Noise filtering considers the use of corpora and prediction technique. The clustering step has two components: cluster POS-variants and POS-homogenous forms.

Given the labeling of each POS variant with multiple features (e.g., type of verb, noun and adjective) at the derivation phase, it is possible to have a more effective POS-variant clustering approach. For instacne, one clustering parameter can be to consider the extent to which the noun and adjective forms are semantically linked with the respective verb form. Table 1 shows that cluster 1 is about someone, while cluster 2 is about an 'object'.

On the other hand, POS-homogenous clustering (see Table 2) allows having intra-POS clusters for nouns, verbs and adjectives of a root. This can be handled using rules from [4]. This illustration is based on examples from the Amharic[15] language.

Table 1. POS-variant cluster example of the root 's-b-r', referring to "break"

Cluster	Verb	Noun	Adjective
1	sebber-e: 'broke, broke-in'	sebar-i: 'one who breaks'	sebber: 'defiant'
2	te-sebber-e: 'was broken'	sIbbar-i: 'fragment'	te-sebbar-i: 'fragile', sebar-a: 'broken'

Table 2. POS-homogenous cluster examples (nouns)

Process: Cluster-1	Object: Cluster-2	State/condition: Cluster-3
seber-a: 'act of breaking'	sebar-a: 'broken piece'	sIbbIrat: 'fracture'
sIbr-iya: 'process of breaking'	sIbbari: 'fragment'	sIbr: 'feeling of strain/hunger'

6 Conclusion and Future Work

In this paper an attempt is made not only to thoroughly survey the literature and justify the need for a novel approach to design derivation lexicon for Semitic languages but also presented illustrated design architecture. In this regard, the concept of FSA in its various forms are relevant in realizing the very early stages of our approach such as in generating surface forms from roots, patterns, vocalism/vowels and rules. We also benefit by using noise filtering strategy. Finally, we also experiment on using a small seed data based word prediction algorithm. However, the later stages such as forming POS-variant derivational clusters and the POS-homogenous clusters require innovate solutions, amounting to important new contributions to the NLP and IR research. It is important to note that the major contribution of this paper is a thorough survey of the literature and illustrated design architecture.

[15] The morphological rules are taken from [4] and English glosses are mainly from Amsalu (1987) and Kane (1990) Amharic-English Dictionaries

References

1. Angelelli, P., Valeria, C., Burani, C.: The Effect of morphology on spelling and reading accuracy: a study on Italian Children. In: Frontiers in Psychology 5(Article No. 1373) (2014).
2. Aqel, A., Alwadei, S., Dahab, Mohammed Y.: Building an Arabic Word Generator. International Journal of Computare Applications 112(14) ,36-41(2015).
3. Attia, M., Pecina, P., Toral, A., Tounsi, L., Genabith, J.V.: An Open-Source Finite State Morphological Transducer for Modern Standard Arabic. In: Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing, pp.125-133. Association for Computational Linguistics, Blois, France (2011).
4. Baye Y.: የአማርኛ ስዋሰው:የተሻሻለ ሶስተኛ እትም. 3rd edn. Addis Ababa University Business Enterprise Printing Press Addis Ababa (2009 E.C[16]).
5. Bender, L, Fullass, H.: Amharic Verb Morphology: A Generative Approach. African Studies Center, Michigan State University, Michigan 48824, USA (1978).
6. Berant, J., Dagan, I., Goldberger, J.: Learning Entailment Relations by Global Graph Structure Optimization. Computational Linguistics 38(1), 73-111 (2012).
7. Bertram, R., Laine, M., Vikkala, M.M.: The Role of Derivational Morphology in Vocabulary Acquisition: Get by with a Little help from my Morpheme Friends. Scandinavian Journal of Psychology 41(4), 287-296(2000).
8. Bezza A.: The Submorphemic Structure of Amharic: Toward a Phono-Semantic Analysis. University of Illinois, USA (2013).
9. Boudelaa, S., Marslen-Wilson, W. D.: Structure, Form, and Meaning in the Mental Lexicon: Evidence from Arabic. Language, Cognition and Neuroscience 30(4), 955-992(2015).
10. Cohen-Sygal, Y., Wintner, S.: Finite-State Registered Automata for Non-Concatenative Morphology. Computational Linguistics 32(1), 49-82(2006).
11. Diab, M., Martin, Y.: Semantic Processing of Semitic Languages. In: Zitouni, I., Hovy, E (eds) Natural Language Processing of Semitic Languages: Theory and Applications of Natural Language Processing, pp. 129-152. Springer, Heidelberg, Berlin, Germany (2014).
12. Freynik, S., Gor, K., O'Rourke, P.: L2 Processing of Arabic Derivational Morphology. The Mental Lexicon 12(1), pp.21-50(2017).
13. Gasser, M.: HornMorpho: A System for Morphological Processing of Amharic, Oromo, and Tigrignya. In: Conference on Human Language Technology for Development, pp. 94-99, Alexanderia, Egypt (2011).
14. Gaussier, É.: Unsupervised Learning of Derivational Morphology from Inflectional Lexicons. In: ACL'99 Workshop Proceedings on Unsupervised Learning in Natural Language Processing, pp.24-30. College Park, Maryland, USA (1999).
15. Habash, N., Rambow, O., Kiraz, G.: Morphological Analysis and Generation for Arabic Dialects. In: Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, pp.17-24. Association for Computational Linguistics, Ann Arbor, USA (2005).
16. Habash, N.: Large Scale Lexeme Based Arabic Morphological Generation. In: JEP-TALN 2004, Session Traitement Automatique de l' Arabe, Fès, Morocco (2004).
17. Habash, N., Dorr, B.: A Categorical Variation Database for English. In: Proceedings of the Annual Meeting of the North American Association for Computational Linguistics, pp.96-102. Edmonton, Canada (2003).

[16] E.C. stands for Ethiopian Calendar, which is 7 years (September to December)/ 8 years (January to August) behind from the Gregorian calendar.

18. Halcomb, T.M.W.: Generating Amharic Present Tense Verbs: A Network Morphology & DATR Account. College of Arts and Sciences, University of Kentucky, MA, USA (2017).
19. Hulden, M.: Foma: A Finite-State Compiler and Library. In: Proceedings of the Demonstration Session at EACL 2009, pp.29-32. Association for Computational Linguistics, Athens (2009a).
20. Hulden, M.: Revisiting Multi-tape Automata for Semitic Morphological Analysis and Generation. In: Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages, pp. 19-26. Association for Computational Linguistics, Athens (2009b).
21. Jurafsky, D., Martin, J. H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. 3rd ed(draft). (2020). https://web.stanford.edu/~jurafsky/slp3/, last accessed 2021/12/03.
22. Kay, M.: Nonconcatenative Finite-State Morphology. In: 3rd Conference of the European Chapter of the Association for Computational Linguistics,pp.2-10. Association for Computational Linguistics, Copenhagen, Denmark (1987).
23. Kibur L. W.: Design and Development of Automatic Morphological Synthesizer for Amharic Perfective Verbs. School of Information Studies for Africa, Addis Ababa University, Addis Ababa (2002).
24. Kiraz, G. A..: Multitiered Nonlinear Morphology Using Multitape Finite Automata: A Case Study on Syriac and Arabic. Computational Linguistics 26(1), 77-105(2000).
25. Koskenniemi, K.: Two-level morphology: A General Computational Model for Word-form Recognition and Production. Department of General Linguistics, University of Helsinki, Hesinki (1983).
26. Levie, R., Ashkenazi, O., Stanzas, S. E., Zwilling, R., Raz, E., Hershkovitz, L., Ravid, D.: The Route to the Derivational Verb Family in Hebrew: A Psycholinguistic Study of Acquisition and Development. Morphology 30, 1–60 (2020).
27. Martinez, A.G.: A Computational Model of Modern Standard Arabic Verbal Morphology based on Generation. Laboratorio, de Linguistica Informatica LLI-UAM, Department de Linguistica, Facultad de Filosofia Letras, Universidad Autonoma de Madrid, Spain (2012).
28. McCarthy, J.: A Prosodic Theory of Non-Concatenative Morphology. Linguistic Inquiry 12(3), 373-417(1981).
29. Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., Grishman, R.: Annotating Noun Argument Structure for NomBank. In: 4th International Conference on Language Resources and Evaluation on Proceedings of LREC 2004, pp. 803-806. European Language Resources Association (ELRA), Lisbon, Portugal (2004).
30. Mohamed, M., Oussalah, M.: A Hybrid Approach for Paraphrase Identification based on Knowledge-Enriched Semantic Heuristics. Language Resources & Evaluation 54,457-485(2020).
31. Moreau, F., Claveau, V., Sébillot, P.: Automatic Morphological Query Expansion Using Analogy-Based Machine Learning. In: Amati, G, Carpineto, C, Romano, G. (eds) Advances in Information Retrieval ECIR 2007, Lecture Notes in Computer Science (LNCS), vol. 4425, pp.222-233. Springer, Heidelberg, Berlin, Germany (2007).
32. Neme, A. Amid.: A lexicon of Arabic verbs constructed on the basis of Semitic taxonomy and using Finite-State Transducers. In: EESSLLI International Workshop on Lexical Resources. WoLeR 2011, Ljubliana, Slovenia (2011).
33. Pad´o, S., Snajder, J, Zeller, B.: Derivational Smoothing for Syntactic Distributional Semantics. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp.731-735. Association for Computational Linguistics, Sofia, Bulgaria (2013).

34. Ravid, D.: Word-Level Morphology: A Psycholinguistic Perspective on Linear Formation in Hebrew Nominals. Morphology 16,127-148(2006).
35. Rodriguez, H., Farwell, D., Farreres, J., Bertran, M., Alkhalifa, M., Marti, M. A., Black, W., Elkateb, S., Kirk, J., Pease, A., Vossen, P., Fellbaum, C.: Arabic WordNet: Current State and Future Extensions. In: Tanács, A., Csendes , D., Vincze ,V., Fellbaum, C., Vossen, P. (Eds.) Proceedings of the Fourth International GlobalWordNet Conference (GWC 2008), pp. 387-405. Department of Informatics, University of Szeged, Szeged, Hungary (2008).
36. Saba A., Gibbon, D. A Complete FS Model for Amharic Morphographemics. In:Yli-Jyra, A., Karttunen, L., Karhumaki, J. (Eds.) FSMNLP 2005, LNAI 4002, pp. 283–284. Springer, Heidelberg (2006).
37. Shaalan, K., Monem, A.A., Rafea, A.: Arabic Morphological Generation from Interlingua: A Rule-based Approach. In: Shi, Z., Shimohara, K., Feng, D. (eds) Intelligent Information Processing III. IIP 2006, IFIP International Federation for Information Processing, vol 228, pp.441-451. Springer, Boston, MA, USA (2006).
38. Sisay F., and Haller, J. Amharic Verb Lexicon in the Context of Machine Translation. In: TALN 2003, Batz-sur-Mer, 11-14 June (2003).
39. Shimron, J.: Semitic languages: Are they really root-based? In: J. Shimron (eds), Language processing and acquisition in languages of Semitic, root-based, morphology, vol. 28, pp.1-28. John Benjamins Publish Company, Amsterdam (2003).
40. Snajder, J.: DERIVBASE.HR: A High-Coverage Derivational Morphology Resource for Croatian. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14), pp.3371-3377. European Language Resources Association (ELRA), Reykjavik, Iceland (2014).
41. Tayler, A., Nagy, W.E.: The Role of Derivational Suffixes in Sentence Comprehension. Technical Report No. 357. Bolt Beranek and Newman Inc, Cambridge, Massachusetts, Boston, USA (1985).
42. Vodolazsky, D.: DeriveBase.Ru: A Derivational Morphology Resource for Russian. In: Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pp. 3937-3943. European Language Resources Association (ELRA), Marseille, France (2020).
43. Wedekind, K.: Which Form Predict all other Best? Variation on the Amharic Verb "Theme". Journal of Ethiopian Studies 25(Nov.1992), 73-92(1992).
44. Zeller, B., Šnajder, J., Pado, S.: DERIVEBASE: Inducing and Evaluating a Derivational Morphology Resource for German. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp.1201-1211.Association for Computational Linguistics, Sofia, Bulgaria(2013).
45. Zitouni, I.: Preface. In: Hovy, E. (eds) Natural Language Processing of Semitic Languages: Theory and Applications of Natural Language Processing, pp. v-viii. Springer, Heidelberg, Berlin, Germany (2014).