A Derivational Chain Bank for Modern Standard Arabic

Reham Marzouk,[1,2] Sondos Krouna,[3] Nizar Habash[1]

[1]Computational Approaches to Modeling Language (CAMeL) Lab,
New York University Abu Dhabi
[2]Information Technology Department, IGSR, Alexandria University
[3]ISLT, Carthage University
marzoukreham@gmail.com, sondes.krouna@islt.ucar.tn, nizar.habash@nyu.edu

Abstract

We introduce the new concept of an *Arabic Derivational Chain Bank* (CHAINBANK) to leverage the relationship between form and meaning in modeling Arabic derivational morphology. We constructed a knowledge graph network of abstract patterns and their derivational relations, and aligned it with the lemmas of the CAMELMORPH morphological analyzer database. This process produced chains of derived words' lemmas linked to their corresponding lemma bases through derivational relations, encompassing 23,333 derivational connections. The CHAINBANK is publicly available.[1]

1 Introduction

Lexical resources are essential for improving the accuracy of language processing and pedagogical applications, as they enhance computational systems' ability to grasp the nuanced meanings and contextual variations of human language. Despite significant efforts, the Arabic language still lacks tools that focus on its compositional morphological structure and semantic connections. Derivational modeling offers a computational framework to capture the interplay between word form and meaning, clarifying Arabic's complex derivational pathways and resolving its structural ambiguities.

Arabic derivational morphology is fundamentally tied to its templatic system, where roots and patterns provide different types of semantic abstractions to express multiple meanings (Gadalla, 2000; Holes, 2004; Habash, 2010). The process of deriving words from roots is not consistent, leading to challenges that hinder the understanding of the meanings of derived words and pose significant obstacles for derivational modeling. For instance, a single pattern can convey different derivational meanings, resulting in ambiguity among derived words that share the same root. As an example, the *masdar/verbal noun* المصدر and the *descriptive adjective* الصفة المشبهة may share the same pattern *1a2A3*, e.g. حصاد *HaSAd*[2] 'harvest' and the adjective جبان *jabAn* 'coward'. Likewise, Homographs can be derived from the same base to convey distinct meanings; consequently, each word possesses a different set of derivatives. For example, each of the two senses of the verb فلح *falaH* 'to succeed' and 'to farm' has its own masdar: فلاح *falAH* 'success' and فلاحة *filAHaħ* 'farming'. Another crucial behavior is the meaning shift of some derivatives from the original abstract meaning of the root, e.g., كتيبة *katiybaħ* 'battalion' is ultimately derived from the root ك.ت.ب *k.t.b* 'writing-related'.

Interpreting the behavior of derived words in the Arabic language, along with the deviations from derivational rules, necessitates a robust organization of derivatives within a framework capable of tracing the various paths of derivation and managing the resulting ambiguities. The objective of this study is to define the new concept of the *Arabic Derivational Chain Bank* (henceforth, CHAINBANK), which serves as the first representation of the Arabic derivational structure. The CHAINBANK presents connected chains that illustrate the path of each derived word and the relation between connected words by providing their derivational meanings. To construct the CHAINBANK, we employed a knowledge graph structure to build a network of abstract patterns, along with a classification model to align this network with lexical database of the Arabic morphological analyzer the CAMELMORPH (Khairallah et al., 2024a) based on selected features. The CHAINBANK is a morphological model that exploits Arabic's compositional morphological and semantic features while accommodating ad hoc exceptions.

[1]https://github.com/CAMeL-Lab/ArabicChainBank

[2]Habash et al. (2007)'s Arabic transliteration scheme.

Processings of the 1st Workshop on NLP for Languages Using Arabic Script (AbjadNLP), pages 78–87
January 19, 2025. ©2025 Association for Computational Linguistics

2 Related Work

Computational Derivational Morphology Several studies have modeled derivational morphology using a range of techniques. Habash and Dorr (2003) clustered categorial variations of English lexemes to develop the CATVAR resource. Similarly, Zeller et al. (2013) created DERIVBASE, a derivational resource for German, using a rule-based framework to induce derivational families (i.e., clusters of lemmas in derivational relationships). Hathout and Namer (2014) developed Démonette by integrating two lexical resources and applying rules to link words to their bases while considering their semantic types. Following Zeller's approach, Vodolazsky (2020) and Šnajder (2014) constructed derivational models for Russian and Croatian, respectively. Kanuparthi et al. (2012) introduced a derivational morphological analyzer for Hindi built on a mapping from an inflectional analyzer. Cotterell et al. (2017) argued for a paradigmatic treatment of derivational morphology and used sequence-to-sequence models to learn mappings from fixed paradigm slots to their corresponding derived forms. Hofmann et al. (2020) proposed a graph auto-encoder that learns embeddings capturing information about the compatibility of affixes and stems in derivation.

Arabic Computational Morphology Research on Arabic computational morphology has primarily focused on inflectional modeling (Kiraz, 1994; Beesley, 1998; Al-Sughaiyer and Al-Kharashi, 2004; Habash and Rambow, 2006; Taji et al., 2018). This focus has led to the development of various models for morphological analysis, generation, and disambiguation. Habash et al. (2012) introduced MADA, a tool designed to analyze and disambiguate Arabic morphology in context. Pasha et al. (2014) developed MADAMIRA, which identifies the morphological features of a word and ranks analysis results based on their compatibility with the model's predictions. More recently, tools such as CALIMA-Star (Taji et al., 2018) and CAMEL-MORPH (Habash et al., 2022; Khairallah et al., 2024b,a) have emerged as advanced morphological analyzers and generators, with a wide range of features. A few efforts have incorporated derivational features to enhance their models. For instance, the morphological analyzer Al Khalil Morph system (Boudlal et al., 2010; Boudchiche et al., 2017) utilizes a database categorized into derived and non-derived classes based on root, vocalized, and unvocalized patterns. Additionaly, Zaghouani et al. (2016) conducted a pilot study aimed at representing the derivational structure of roots and patterns while addressing the multiple senses associated with a single pattern. However, none of these studies developed a comprehensive model focused extensively on derivational morphology.

Inspired by the efforts on systematic treatment of derivational morphology in other languages, we propose a model that captures the complexity and elegance of the Arabic derivational system.

3 Arabic Derivational Morphology Terms

In Arabic templatic morphology, discontinuous consonantal morphemes, **roots**, interconnect with different **patterns** of vowels and consonants to construct different meanings. Each root has a general semantic meaning and each pattern is associated with a certain **canonical meaning**. The set of words sharing the same root, a **derivational family**, are organizable as a derivational network connecting hierarchically up to a (typically) single base word. Derived words can be either **canonical**, where the word's meaning matches its pattern's meaning, or **non-canonical**, where an ad hoc deviation of regular form occurs. For example the two words ضرب *Darb* 'hitting', and شمس *šams* 'sun', share the same pattern *1a23*, whose canonical meaning is the masdar, matching the former (canonical) but not the latter (non-canonical). Derived words can also be formed with **derivational affixes**, e.g., the suffix +ي *+iy∼* (ياء النسبة Attributive yA') appends to the base علم *ςilm* 'science' to produce the attributive adjectives علمي *ςilmiy∼* 'scientific'. Verbs are divided into **unaugmented**, which are composed of roots and vocalism-only patterns, and **augmented**, which are derived from unaugmented verbs by geminating, lengthening of vowels, prefixation or infixation (Gadalla, 2000). Nouns are categorized into **primary nouns**, which are directly derived from roots (Gadalla, 2000), and **derived nouns**, which originate from verbs and encompass derivational classes such as verbal nouns (masdar), nouns of location, etc. In some cases, derived words involve shifting the meaning to a contextually unrelated interpretation of their base form, i.e., **semantic specification**. For instance, the noun مكتوب *maktuwb* 'letter/message' is derived from the passive participle مكتوب *maktuwb* 'written".

4 The ChainBank Framework

The role of the Arabic derivational ChainBank is to systematically link all derivatives belonging to the same derivational family in a sequential manner (chain), starting from the root and progressing through each derived form. This chain establishes a clear relationship between each derivative and its base, clarifying the morphological processes that generate new words. By organizing derivatives in this structured form, the derivational chain highlights the hierarchical and interdependent nature of word formation, providing insights into how base forms evolve into more complex derivatives while preserving their semantic and grammatical connections.

We represent the ChainBank as a dynamic tree-structured knowledge graph starting with the root. Each node in this graph corresponds to a derived word and includes its morphosemantic attributes, such as pattern, part-of-speech, functional features, and lexical meaning. The connections between pairs of nodes denote the derivational relationships of each child node to its base parent.

To create the ChainBank, we developed an extensive network that represents the organization of abstract patterns, such as فَعَل *CaCaC/1a2a3* and فَعِيل *CaCiyC/1a2iy3*, and integrated this network with the CamelMorph lexical database. This combination forms a large-scale network connecting Arabic words through their derivational relationships. The process includes two levels:

- The **abstract level** focuses on the abstract patterns designed to represent various derivatives.
- The **concrete level** is where abstract patterns are linked to lemmas to produce derived words along with their derivational meanings.

4.1 The Abstract Level

The network we developed covers all potential connections between roots and their derived patterns in a tree structure. The roots are positioned at the apex of the tree, followed by unaugmented verbs, and subsequently the augmented verbs along with the nominal derivatives. This network is meticulously organized to display all conceivable connections between patterns, even if certain connections may not be attested but remain theoretically plausible.

Constructing the Abstract Network First, we classified all patterns according to their morphosemantic characteristics. Appendix A (Table 2) presents examples of the adopted classification of the patterns selected in this study.

Next, we devised a scheme to incorporate the derivational features of these patterns into the network. The construction of the network involved the establishment of three tables to represent the source and target nodes, along with their relationships.

1. **The Canonical Table** We manually constructed a table that covers trilateral and quadrilateral verbs in their unaugmented and augmented forms as well as all their derived nominal patterns. We present examples of the Canonical Table entries in Appendix C: Table 4 focuses on forms connected to the triliteral unaugmented verbs (Form I, فعل *CaCaC/1a2a3*), and Table 5 includes the rest.
2. **The Affixational Table** To model affixational derivatives, we automatically generate new entries combining Canonical Table entries with specific derivational affixes. For instance, the Canonical Table pattern *1i23* is extended to *1i23+iy∼* to allows us to model the example علمي *ςilmiy∼* 'scientific' from علم *ςilm* 'science' discussed in Section 3.
3. **The Semantic Specification Table** To model derivations that involve a semantic specification shift without a change in the main abstract pattern, we automatically generate entries from the Canonical and Affixational Tables, under a set of manually specified constraints. A major type of such entries involves (but not only) a change in part-of-speech; and in some cases, it may use an inflected form such as the feminine singular or broken plural. For instance, the derivation *ma12uw3+aħ* (noun) from *ma12uw3* (adjective) allows us to model the derivation of معلومة *maςluwmaħ* 'a piece of information (noun)' from the feminine form of معلوم *maςluwm* 'known (adj)'.

The final Abstract Network is built as a combination of the above-mentioned tables in one relational database to allow for efficient access to the chains of connected patterns, all their features, and their derivational relations.

4.2 The Concrete Level

Next we discuss aligning the CamelMorph database lemmas with the abstract network to create the ChainBank.

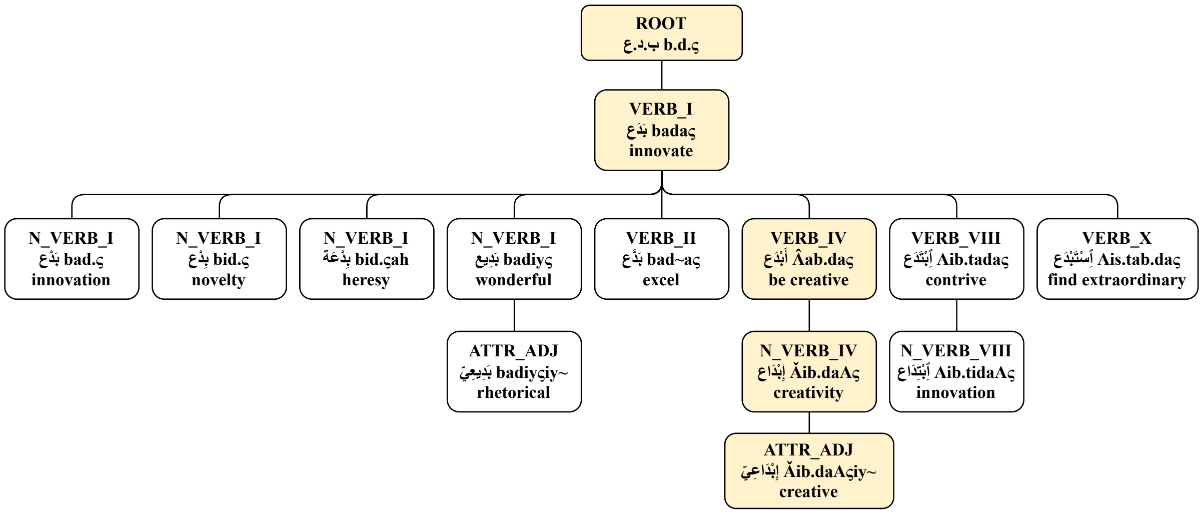

Figure 1: An example of a collection of derivative chains from one root. Highlighted is a chain that links a number of lemmas in derivational progression: the root ب.د.ع *b.d.ς* 'innovation related' ⇒ the verb بَدَع *badaς* 'to innovate' ⇒ the verb أَبْدَع *Âb.daς* 'to be creative' ⇒ the noun إِبْدَاع *ǍibdaAς* 'creativity' ⇒ the adjective إِبْدَاعِيّ *ǍibdaAςiy∼* 'creative'.

For each collection of lemmas from the same root, a derivational family, we recursively construct a tree starting with the root. We only add children (derived lemmas) to parents (derivational bases or the root) in the tree if they match an allowable abstract network pair in terms of all linguistic attributes of child and parent.[3] If a lemma could be paired with different parents or with the same parent but with different relations we duplicate the child lemma and assign it as many times as needed. We continue to assign children to parents until we exhaust all possible pairings. The result is ideally a fully connected tree (knowledge graph) starting with the root of the derivational family and including chains that link it to every lemma in the family. In addition to the lemmas, the nodes of the tree include key linguistic attributes such as the part-of-speech and derivational class. Figure 1 is an example from the CHAINBANK.

In some cases, we may end up with disconnected subtrees due to a lack of allowed pairings in the abstract network. This may be the result of patterns that are disused with some roots.[4]

[3]In some cases, we require an inflectional process as an intermediary stage to produce a new derivation pattern, e.g., deriving a lemma pattern from the plural form of its base lemma: the attributive adjective حدودي *Huduwdiy∼* 'bordering' is derived from the plural form of حد *Had∼* 'border' (حدود *Huduwd* 'borders').

[4]A solution to consider in the future is to force attach such disconnected subtrees to the root with an *Unknown* relation.

5 Evaluation

5.1 Experimental Setup

Gold Chains We manually constructed a set of 100 CHAINBANK trees correspondign to 100 randomly selected roots from CAMELMORPH. To speed up the process, we started with automatically generated trees using an earlier version of our approach, and manually corrected them.

Data Splits We split our 100 CHAINBANK trees into two sets: Dev (25 roots) and Test (75 roots). We used the Dev set to help debug and improve the tables we created for the abstract network and optimize our algorithm. The Test is used for reporting on the final implementation.

Metrics we report on the following metrics.

- Detected Relations (%) is the percentage of all Gold Chain lemmas we detected automatically.
- Single Relation Correct (%) is the percentage of all Gold Chain detected relations that unambiguous and correct.
- Multiple Relation Correct (%) is the percentage of all Gold Chain detected relations that ambiguous and but with one correct answer.
- No Correct Relations (%) is the percentage of remaining detected relations.

	Assessment 1: Dev		Assessment 2: Test		All	
Roots	25		75		4,924	
Lemmas	566		1,608		34,453	
Detected Relations	496	(87.63%)	1,147	(71.33%)	23,333	(67.72%)
Single Relation Correct	450	(90.73%)	1,058	(92.24%)		
Multiple Relation Correct	45	(9.07%)	76	(6.63%)		
No Correct Relation	1	(0.20%)	13	(1.13%)		

Table 1: Results of constructing the relational derivational CHAINBANK using the CAMELMORPH database, on development (Dev) and test subsets of the roots, and on all roots.

5.2 Results and Discussion

The results on Dev show a high degree of detected relations, but not perfect (∼88%), with over 90% single relation correct. The Test is lower in terms of detected relations (∼71%), but a higher single relation correct (92%). Multiple relations, accounting for ∼6-9% of the cases in Dev and Test, occur due to shared patterns across different derivational classes. Missing relations stem from three main factors: (i) the relational data lacks primary nouns and other nominal lemmas, which require specific paths in the CHAINBANKs; (ii) CAMELMORPH's database wasn't designed for derivational modeling, resulting in incomplete lemma groups for some roots and chain disconnections; or (iii) the relational database needs expansion with new non-canonical patterns. Additionally, the system could be improved by adding features and techniques to resolve ambiguities during evaluation.

All results are presented in Table 1.

5.3 CHAINBANK v1.0

We further applied our system on 4,926 roots from CAMELMORPH and their lemmas, which resulted in 23,333 relations (∼68% detected relations), constituting the first version of the CHAINBANK. We plan to manually correct and further annotate additional entries in the future. The Arabic Derivational CHAINBANK v1.0 is publicly available to support further Arabic NLP research.[5]

6 Conclusion and Future Work

We introduced the Arabic Derivational CHAINBANK framework for modeling Arabic derivational morphology. The evaluation of our rule-based method to populate the CHAINBANK shows great promise. The first edition of the CHAINBANK and its framework are publicly available.

Future work will continue to expand the abstract network to include missing patterns, including non-canonical patterns, and to develop advanced disambiguation techniques to further enhance the CHAINBANK. This work should happen in tandem with improving the coverage of CAMELMORPH in terms of lemmas and their features. Our long term vision is to include dialectal lemmas in a manner that shows their connections with each other and with Standard Arabic lemmas.

7 Limitations

We acknowledge several limitations of the work as presented. First, the reliance on a rule-based methodology, although efficient, may overlook nuances that a more comprehensive manual annotation process could capture. This could lead to the omission of certain derivational patterns and relations. Second, the alignment with the CAMELMORPH morphological analyzer, though beneficial for broad coverage, may have resulted in incomplete or fragmented derivational chains due to the database's current structure, which was not designed for derivational modeling. Third, the dataset predominantly covers canonical derivational patterns, with non-canonical patterns remaining underrepresented, potentially limiting the CHAINBANK's applicability to broader linguistic phenomena. Lastly, the work focuses on Standard Arabic and does not cover any of its major dialects. Future work will address these limitations to enhance the framework's completeness and accuracy.

References

Imad A. Al-Sughaiyer and Ibrahim A. Al-Kharashi. 2004. Arabic morphological analysis techniques: A comprehensive survey. *Journal of the American Society for Information Science and Technology*, 55(3):189–213.

Kenneth Beesley. 1998. Arabic morphology using only

[5] https://github.com/CAMeL-Lab/ArabicChainBank

finite-state operations. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages (CASL)*, pages 50–7, Montereal.

Mohamed Boudchiche, Azzeddine Mazroui, Mohamed Ould Abdallahi Ould Bebah, Abdelhak Lakhouaja, and Abderrahim Boudlal. 2017. AlKhalil Morpho Sys2: A robust Arabic morpho-syntactic analyzer. *Journal of King Saud University-Computer and Information Sciences*, 29(2):141–146.

Abderrahim Boudlal, Abdelhak Lakhouaja, Azzeddine Mazroui, Abdelouafi Meziane, MOAO Bebah, and Mostafa Shoul. 2010. AlKhalil Morpho Sys1: A morphosyntactic analysis system for Arabic texts. In *International Arab conference on information technology*, pages 1–6.

Ryan Cotterell, Ekaterina Vylomova, Huda Khayrallah, Christo Kirov, and David Yarowsky. 2017. Paradigm completion for derivational morphology. *arXiv preprint arXiv:1708.09151*.

Hassan Gadalla. 2000. *Comparative Morphology of Standard and Egyptian Arabic*. LINCOM EUROPA.

Nizar Habash and Bonnie Dorr. 2003. CatVar: a database of categorial variations for English. In *Proceedings of Machine Translation Summit IX: System Presentations*, New Orleans, USA.

Nizar Habash, Reham Marzouk, Christian Khairallah, and Salam Khalifa. 2022. Morphotactic modeling in an open-source multi-dialectal arabic morphological analyzer and generator. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 92–102.

Nizar Habash and Owen Rambow. 2006. MAGEAD: A morphological analyzer and generator for the Arabic dialects. In *Proceedings of the International Conference on Computational Linguistics and the Conference of the Association for Computational Linguistics (COLING-ACL)*, pages 681–688, Sydney, Australia.

Nizar Habash, Owen Rambow, and Ryan Roth. 2012. MADA+TOKAN Manual. Technical report, Technical Report CCLS-12-01, Columbia University.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pages 15–22. Springer, Netherlands.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

Nabil Hathout and Fiammetta Namer. 2014. Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology*, 11.

Valentin Hofmann, Hinrich Schütze, and Janet Pierrehumbert. 2020. A graph auto-encoder model of derivational morphology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1127–1138, Online. Association for Computational Linguistics.

Clive Holes. 2004. *Modern Arabic: Structures, functions, and varieties*. Georgetown University Press.

Nikhil Kanuparthi, Abhilash Inumella, and Dipti Misra Sharma. 2012. Hindi derivational morphological analyzer. In *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*, pages 10–16.

Christian Khairallah, Salam Khalifa, Reham Marzouk, Mayar Nassar, and Nizar Habash. 2024a. Camel morph MSA: A large-scale open-source morphological analyzer for Modern Standard Arabic. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2683–2691, Torino, Italia. ELRA and ICCL.

Christian Khairallah, Reham Marzouk, Salam Khalifa, Mayar Nassar, and Nizar Habash. 2024b. Computational morphology and lexicography modeling of Modern Standard Arabic nominals. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1071–1084, St. Julian's, Malta. Association for Computational Linguistics.

George Kiraz. 1994. Multi-tape Two-level Morphology: A Case study in Semitic Non-Linear Morphology. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 180–186, Kyoto, Japan.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1094–1101, Reykjavik, Iceland.

Jan Šnajder. 2014. DerivBase.hr: A high-coverage derivational morphology resource for Croatian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3371–3377, Reykjavik, Iceland. European Language Resources Association (ELRA).

Dima Taji, Jamila El Gizuli, and Nizar Habash. 2018. An Arabic dependency treebank in the travel domain. In *Proceedings of the Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT)*, Miyazaki, Japan.

Daniil Vodolazsky. 2020. Derivbase. ru: A derivational morphology resource for Russian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3937–3943.

Wajdi Zaghouani, Abdelati Hawwari, Mona Diab, Tim O'Gorman, and Ahmed Badran. 2016. AMPN: a semantic resource for Arabic morphological patterns. *International Journal of Speech Technology*, 19:281–288.

Britta Zeller, Jan Šnajder, and Sebastian Padó. 2013. DErivBase: Inducing and evaluating a derivational morphology resource for German. In *Proceedings of the Association for Computational Linguistics*, pages 1201–1211.

A CHAINBANK Derivational Classes

	ChainBank Derivational Classes	POS	Morphosemantic Feature	Example: Pattern	Example: Arabic	Example: Gloss
(1)	Masdar المصدر	Noun	**Event حدث**	ta1a2~u3,..	تعلُّم	learning
(2)	M+Masdar المصدر الميمي	Noun		ma1o2a3	مسمع	hearing
(3)	Noun of Masdar اسم المصدر	Noun		ta1A2u3+iy~ap	تفاعلية	interaction
(4)	Masdar+iy~a المصدر الصناعي	Noun	**Entity الذات**	1a2iy3+ap	غنيمة	loot
(5)	Noun اسم جنس	Noun		1a2a3	جبل	mountain
(6)	Noun of Location اسم المكان	Noun		ma1o2a3	مخرَج	exit
(7)	Noun of Instrument اسم الآلة	Noun		مِفعال ومِفعلة/mi1o2a3	منحت	chisel
(8)	Noun of Instance اسم المرة	Noun		1a2o3+ap	أكلة	meal
(9)	Semantic Specification التخصيص الدلالي	Noun		ma1o2uw3+ap	معلومة	information
(10)	Descriptive Adjective الصفة المشبهة	Adj	**Consistency ثبوت**	1a2uw3,....	خجول	shy
(11)	Comparative Adjective اسم التفضيل	Adj		>a1o2a3	أعقل	wiser
(12)	Attributive Adjective اسم النسبة	Adj		1a2a3+iy~, ...	عمليّ	practical
(13)	Active Participle اسم الفاعل	Adj	**Changeability تجدد**	1A2i3, ...	آكل	eater
(14)	Passive Participle اسم المفعول	Adj		ma1o2uw3,...	مسموع	being heard
(15)	Noun of Exaggeration اسم المبالغة	Adj		1a2iy3, ...	شريب	drunkard
(16)	Form_I فَعَل	Verb	Basic root meaning (T/I)	1a2a3	أكل	to eat
(17)	Form_I فعِل	Verb	Stative(T/I)	1a2i3	شرب	to hear
(18)	Form_I فعُل	Verb	Attributing an adjective (I)	1a2u3	حسُن	to be beautiful
(19)	Form_II فَعَّل	Verb	Transitivity (T/I)	1a2~a3	قوّى	to stregnthen
(20)	Form_III فاعل	Verb	Reciprocality (T/I)	1A2a3	ذاكر	to memorize
(21)	Form_IV أفعَل	Verb	Causative/Factitive (T/I)	>a1o2a3	أخرج	to get out
(22)	Form_V تفعَّل	Verb	Intransitivity/ Compliance (T/I)	ta1a2~a3	تعلّم	to learn
(23)	Form_VI تفاعل	Verb	Reciprocal /Compliance (T/I)	ta1A2a3	تفاعل	to interact
(24)	Form_VII انفعل	Verb	Intransitivity/ Compliance (I)	{in1a2a3	انجذب	to be attracted
(25)	Form_VIII افتعل	Verb	Reciprocality/Intensivity(T/I)	{i1ota2a3	اغتنم	to gain
(26)	Form_IX افعلّ	Verb	Colors / Defects (T/I)	{i1o2a3~	احمرّ	to get red
(27)	Form_X استفعل	Verb	Doing an action(T/I)	{isota1o2a3	استخرج	to extract
(28)	QUAD_Form_I فعلل	Verb	Doing an action (T/I)	1a2o3a4	زخرف	to decorate
(29)	QUAD_Form_II افعللّ	Verb	Intensivity(I)	{i1o2a3a4~	اطمأن	to reassure
(30)	QUAD_Form_III تفعلل	Verb	Intransitivity/ Compliance (I)	ta1a2o3a4	تقهقر	to retreat

Table 2: CHAINBANK Derivational Classes: an overview of Arabic Morphosemantic patterns with examples. (T/I) refers to transitive & intransitive

B CHAINBANK Tags

ID	TAG	Derivational Class	POS	LOCATION
1	ROOT	Triconsonantal Root	ROOT	Canonical_TABLE
2	V_I	Verb:form I	VERB	Canonical_TABLE
3	V_*	Verb:form *	VERB	Canonical_TABLE
4	QUAD_V_I	Quadriliteral Verb	VERB	Canonical_TABLE
5	QUAD_V_*	Quadriliteral Verb	VERB	Canonical_TABLE
6	N_VERB_I	Masdar:I	NOUN	Canonical_TABLE
7	N_VERB_*	Masdar:*	NOUN	Canonical_TABLE
8	N_QUAD_VERB_I	Quadriliteral Masdar:I	NOUN	Canonical_TABLE
9	N_QUAD_VERB_*	Quadriliteral Masdar:*	NOUN	Canonical_TABLE
10	N_LOC	Noun of Location	NOUN	Canonical_TABLE
11	N_INSTR	Noun of Instrument	NOUN	Canonical_TABLE
12	N_MANN	Noun of Manner	NOUN	Canonical_TABLE
13	N_INST	Noun of Instance	NOUN	Canonical_TABLE
14	M_MAS	M+Masdar	NOUN	Canonical_TABLE
15	COMP_ADJ	Comparative Adjective	ADJ	Canonical_TABLE
16	N_EXAG	Noun of Exaggeration	ADJ	Canonical_TABLE
17	ACT_PRTC	Active Participle	ADJ	Canonical_TABLE
18	PASS_PRTC	Passive Participle	ADJ	Canonical_TABLE
19	DES_ADJ	Descriptive Adjective	ADJ	Canonical_TABLE
20	QUAD_ACT_PRTC	Quadriliteral Active Participle	ADJ	Canonical_TABLE
21	QUAD_PASS_PRTC	Quadriliteral Passive Participle	ADJ	Canonical_TABLE
22	QUAD_ACT_PRTC	Quadriliteral Active Participle	ADJ	Canonical_TABLE
23	QUAD_PASS_PRTC	Quadriliteral Passive Participle	ADJ	Canonical_TABLE
24	ADJ_ATTR	Attributive Adjective	ADJ	Affixation_TABLE
25	MAS_iy~a	Masdar+iy~a	NOUN	Affixation_TABLE
26	SEM_SPECS	Semantic Specification	NOUN	Semantic_Specs_TABLE

Table 3: Derivational Classes in the Arabic CHAINBANK, showing ID, tag, class, POS, and location of each form within the derivational network.

C ChainBank Canonical Tables

PARENT:PATTERN	PARENT_POS	CHILD_DER_CAT	CHILD:PATTERN	CHILD_POS	CHILD_EXAMPLE
None	NONE	ROOT	ROOT	ROOT	"ك.ت.ب"
ROOT	ROOT	VERB_I	V_I:1a2a3	VERB	أكل
V_I:1a2a3	VERB	PASS_PRTC	PASS_PRTC:ma1o2uw3	ADJ	مكتوب
V_I:1a2a3	VERB	N_LOC	N_LOC:ma1o2i3+ap	NOUN	منطقة
V_I:1a2a3	VERB	N_LOC	N_LOC:ma1o2i3	NOUN	مجلس
V_I:1a2a3	VERB	N_LOC	N_LOC:ma1o2a3+ap	NOUN	مكتبة
V_I:1a2a3	VERB	N_INSTR	N_INSTR:mi1o2a3	NOUN	مبرد
V_I:1a2a3	VERB	N_INSTR	N_INSTR:1a2~A3+ap	NOUN	سماعة
V_I:1a2a3	VERB	N_INST	N_INST:1a2o3+ap	NOUN	ضربة
ACT_PRTC:1A2i3	ADJ	N_EXAG	N_EXAG:1a2iy3	ADJ	حليم
ACT_PRTC:1A2i3	ADJ	N_EXAG	N_EXAG:1a2~A3	ADJ	علام
V_I:1a2a3	VERB	N_VERB_I	MASDAR:1u2uw3+ap	NOUN	برودة
V_I:1a2a3	VERB	N_VERB_I	MASDAR:1u2uw3	NOUN	دخول
V_I:1a2a3	VERB	N_VERB_I	MASDAR:1u2o3An	NOUN	غفران
V_I:1a2a3	VERB	N_VERB_I	MASDAR:1i2A3+ap	NOUN	كتابة
V_I:1a2a3	VERB	N_VERB_I	MASDAR:1i2A3	NOUN	قيام
V_I:1a2a3	VERB	N_VERB_I	MASDAR:1a2uw3	NOUN	قبول
V_I:1a2a3	VERB	N_VERB_I	MASDAR:1a2o3aY	NOUN	شكوى
V_I:1a2a3	VERB	N_VERB_I	MASDAR:1a2o3	NOUN	ضرب
V_I:1a2a3	VERB	N_VERB_I	MASDAR:1a2iy3	NOUN	رحيل
V_I:1a2a3	VERB	N_VERB_I	MASDAR:1a2A3+ap	NOUN	ضخامة
V_I:1a2a3	VERB	N_VERB_I	MASDAR:1a2a3	NOUN	طلب
V_I:1a2a3	VERB	N_VERB_I	MASDAR:1a2A3	NOUN	فساد
V_I:1a2a3	VERB	MASDAR_M	MASDAR_M:ma1o2i3+ap	NOUN	معرفة
V_I:1a2a3	VERB	MASDAR_M	MASDAR_M:ma1o2a3	NOUN	مقتل
V_I:1a2a3	VERB	MASDAR_M	MASDAR_M:ma1iy3	NOUN	مصير
V_I:1a2a3	VERB	MASDAR_M	MASDAR_M:ma1a3o3+ap	NOUN	مودة
V_I:1a2a3	VERB	MASDAR_M	MASDAR_M:ma1A3	NOUN	منام
V_I:1a2a3	VERB	DES_ADJ	ADJ:ma1o2uw3	ADJ	موفور
V_I:1a2a3	VERB	DES_ADJ	ADJ:1u2u3	ADJ	صعب
V_I:1a2a3	VERB	DES_ADJ	ADJ:1u2o3An	ADJ	عريان
V_I:1a2a3	VERB	DES_ADJ	ADJ:1u2o3	ADJ	حلو
V_I:1a2a3	VERB	DES_ADJ	ADJ:1u2A3	ADJ	شجاع
V_I:1a2a3	VERB	DES_ADJ	ADJ:1a2uw3	ADJ	وقور
V_I:1a2a3	VERB	DES_ADJ	ADJ:1a2iy3	ADJ	عظيم
V_I:1a2a3	VERB	DES_ADJ	ADJ:1a2o3	ADJ	ضخم
V_I:1a2a3	VERB	DES_ADJ	ADJ:1a2~i3	ADJ	طيب
V_I:1a2a3	VERB	DES_ADJ	ADJ_M:1a2o3An	ADJ	عطشان
V_I:1a2a3	VERB	DES_ADJ	ADJ_M:>a12a3	ADJ	أسمر
V_I:1a2a3	VERB	DES_ADJ	ADJ_F:1a2O3aY	ADJ	عطشى
V_I:1a2a3	VERB	DES_ADJ	ADJ_F:1a23A'	ADJ	سمراء
ADJ:1a2uw3	ADJ	COMP_ADJ	ADJ_COMP_9	ADJ_COMP	أخجل
ADJ:1a2iy3	ADJ	COMP_ADJ	ADJ_COMP_6	ADJ_COMP	أرشق

Table 4: (Part I) A sample of entries from `Canonical_I` table: one of the fundamental tables in the ChainBank.

PARENT:PATTERN	PARENT_POS	CHILD_DER_CAT	CHILD:PATTERN	CHILD_POS	CHILD_EXAMPLE
ROOT	VERB	QUAD_VERB_I	QUAD_V_I:1a2o3a4	VERB	زخرف
V_I:1a2a3	VERB	VERB_II	V_II:1a2~a3	VERB	كسَّر
V_I:1a2i3	VERB	VERB_II	V_II:1a2~a3	VERB	كبَّر
V_I:1a2u3	VERB	VERB_II	V_II:1a2~a3	VERB	كبَّر
V_I:1a2a3	VERB	VERB_III	V_III:1A2a3	VERB	سافر
V_I:1a2a3	VERB	VERB_IV	V_IV:>a1o2a3	VERB	أوجد
V_I:1a2i3	VERB	VERB_IV	V_IV:>a1o2a3	VERB	أخرج
V_I:1a2i3	VERB	VERB_VIII	V_VIII:{i1ota2a3	VERB	اخترق
V_I:1a2u3	VERB	VERB_VIII	V_VIII:{i1ota2a3	VERB	اختبر
V_I:1a2a3	VERB	VERB_VII	V_VII:{ino1a2a3	VERB	اندرج
V_I:1a2a3	VERB	VERB_IX	V_IX:{i1o2a3~	VERB	اخضر
V_I:1a2a3	VERB	VERB_X	V_X:{isota1o2a3	VERB	استحضر
V_II:1a2~a3	VERB	N_VERB_II	MASDAR:ti1o2A3	NOUN	تكرار
V_II:1a2~a3	VERB	N_VERB_II	MASDAR:ta1o2iy3	NOUN	تكبير
V_II:1a2~a3	VERB	N_VERB_II	MASDAR:ta1o2i3+ap	NOUN	تفرقة
V_III:1A2a3	VERB	N_VERB_III	MASDAR:mu1A2a3+ap	NOUN	محاسبة
V_III:1A2a3	VERB	N_VERB_III	MASDAR:1i2A3	NOUN	حساب
V_VI:ta1A2a3	VERB	N_VERB_VI	MASDAR:ta1A2u3	NOUN	تحول
V_VIII:{i1ota2a3	VERB	N_VERB_VIII	MASDAR:{i1oti2A3	NOUN	اختراق
V_IX:{i1o2a3~	VERB	N_VERB_IX	MASDAR:{i1o2i3A3	NOUN	اخضرار
V_X:{isota1o2a3	VERB	N_VERB_X	MASDAR:{isoti1o2A3	NOUN	استحضار
QUAD_V_I:1a2o3a4	VERB	N_QUAD_VERB_I	QUAD_MASDAR:1a2o3a4+ap	NOUN	زخرفة
QUAD_V_II:ta1a2o3	VERB	N_QUAD_VERB_II	QUAD_MASDAR:ta1a2o3u4	NOUN	تدهوُر
V_III:1A2a3	VERB	MASDAR_M	MASDAR_M:mu1A2a3	NOUN	مفاعل
V_IV:>a1o2a3	VERB	MASDAR_M	MASDAR_M:ma1o2a3	NOUN	محضر
V_II:1a2~a3	VERB	DES_ADJ	ADJ:mu1a2~i3	ADJ	معمر
V_III:1A2a3	VERB	DES_ADJ	ADJ:mu1o2i3	ADJ	مؤمن
V_III:1A2a3	VERB	DES_ADJ	ADJ:mu1A2i3	ADJ	مغامر
V_II:1a2~a3	VERB	ACT_PRTC	ACT_PRTC:mu1a2~i3	ADJ	محطِّم
V_III:1A2a3	VERB	ACT_PRTC	ACT_PRTC:mu1A2i3	ADJ	محاصِر
V_IV:>a1o2a3	VERB	ACT_PRTC	ACT_PRTC:mu1o2i3	ADJ	مخبر
V_V:ta1a2~a3	VERB	ACT_PRTC	ACT_PRTC:muta1a2~i3	ADJ	متفهم
V_VI:ta1A2a3	VERB	ACT_PRTC	ACT_PRTC:muta1A2i3	ADJ	متبادل
V_VII:{ino1a2a3	VERB	ACT_PRTC	ACT_PRTC:muno1a2i3	ADJ	مندرِج
QUAD_V_IV:{i1o2a3	VERB	ACT_PRTC	QUAD_ACT_PRTC:mu1o2a3i4	ADJ	مطمئِن
QUAD_V_I:1a2o3a4	VERB	ACT_PRTC	QUAD_ACT_PRTC:mu1a2o3i4	ADJ	مزخرِف
V_II:1a2~a3	VERB	PASS_PRTC	PASS_PRTC:mu1a2~a3	ADJ	معطَّل
V_III:1A2a3	VERB	PASS_PRTC	PASS_PRTC:mu1A2a3	ADJ	محاصَر
V_IV:>a1o2a3	VERB	PASS_PRTC	PASS_PRTC:mu1o2a3	ADJ	مرغم
V_V:ta1a2~a3	VERB	PASS_PRTC	PASS_PRTC:muta1a2~a3	ADJ	متأسلم
V_VI:ta1A2a4	VERB	PASS_PRTC	PASS_PRTC:muta1A2a3	ADJ	متآكل
V_X:{isota1o2a3	VERB	PASS_PRTC	PASS_PRTC:musota1o2a3	ADJ	مستحضَر

Table 5: (Part II) A sample of entries from `Canonical_*` table: one of the fundamental tables in the CHAINBANK.