

Functional Arabic Morphology

Formal System and Implementation

Otakar Smrž

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague

Prague 2007

He will notify them about that through SMS messages, the Internet, and other means. سَيُخْبِرُهُمْ بِذَلِكَ عَنِ طَرِيقِ الرِّسَائِلِ الْقَصِيرَةِ وَالْإِنْتَرِنْتِ وَغَيْرِهَا.

String	Token	Token Tag	Buckwalter's M-Tags	Token Form	Token Gloss
		F-----	FUT	<i>sa-</i>	will
سيخبرهم		VIIA-3MS--	IV3MS+IV+IVSUFF_MOOD:I	<i>yu-ḥbir-u</i>	he-notify
		S----3MP4-	IVSUFF_DO:3MP	<i>-hum</i>	them
بذلك		P-----	PREP	<i>bi-</i>	about/by
		SD----MS--	DEM_PRON_MS	<i>dālika</i>	that
عن		P-----	PREP	<i>ʿan</i>	by/about
طريق		N-----2R	NOUN+CASE_DEF_GEN	<i>ṭarīq-i</i>	way-of
الرسائل		N-----2D	DET+NOUN+CASE_DEF_GEN	<i>ar-rasāʾil-i</i>	the-messages
القصيرة		A----FS2D	DET+ADJ+NSUFF_FEM_SG+ +CASE_DEF_GEN	<i>al-qaṣīr-at-i</i>	the-short
والإنترنت		C-----	CONJ	<i>wa-</i>	and
		Z-----2D	DET+NOUN_PROP+ +CASE_DEF_GEN	<i>al-ʾinternet-i</i>	the-internet
وغيرها		C-----	CONJ	<i>wa-</i>	and
		FN-----2R	NEG_PART+CASE_DEF_GEN	<i>ḡayr-i</i>	other/not-of
		S----3FS2-	POSS_PRON_3FS	<i>-hā</i>	them

Functional Arabic Morphology

Many computational models of Arabic morphology are **lexical** in nature. As they are not designed in connection with any **syntax–morphology interface**, their interpretation is destined to be **incremental**.

Functional Arabic Morphology

Many computational models of Arabic morphology are **lexical** in nature. As they are not designed in connection with any **syntax–morphology interface**, their interpretation is destined to be **incremental**.

Functional Arabic Morphology endorses the **inferential–realizational** views.

Functional Arabic Morphology

Many computational models of Arabic morphology are **lexical** in nature. As they are not designed in connection with any **syntax–morphology interface**, their interpretation is destined to be **incremental**.

Functional Arabic Morphology endorses the **inferential–realizational** views.

It re-establishes the **system** of **inflectional** and **inherent** morphosyntactic properties and distinguishes precisely the **senses** of their use in the grammar.

Functional Arabic Morphology

Many computational models of Arabic morphology are **lexical** in nature. As they are not designed in connection with any **syntax–morphology interface**, their interpretation is destined to be **incremental**.

Functional Arabic Morphology endorses the **inferential–realizational** views.

It re-establishes the **system** of **inflectional** and **inherent** morphosyntactic properties and distinguishes precisely the **senses** of their use in the grammar.

Definition of **lexemes** can include the derivational **root and pattern** information if appropriate. Modeling of the **written** language as well as **spoken** dialects is expected to be methodologically **identical**.

ElixirFM

ElixirFM is a high-level implementation of **Functional Arabic Morphology**.

ElixirFM

ElixirFM is a high-level implementation of **Functional Arabic Morphology**.

ElixirFM uses the Functional Morphology library for **Haskell** and extends it.

ElixirFM

ElixirFM is a high-level implementation of Functional Arabic Morphology.

ElixirFM uses the Functional Morphology library for Haskell and extends it.

Morphology is modeled in terms of paradigms, grammatical categories, lexemes and word classes. The computation of analysis or generation is conceptually distinguished from the general-purpose linguistic model.

ElixirFM

ElixirFM is a high-level implementation of **Functional Arabic Morphology**.

ElixirFM uses the Functional Morphology library for **Haskell** and extends it.

Morphology is **modeled** in terms of **paradigms**, grammatical **categories**, **lexemes** and word **classes**. The **computation** of analysis or generation is conceptually **distinguished** from the **general-purpose** linguistic **model**.

The lexicon of ElixirFM is derived from the open-source **Buckwalter lexicon**, **redesigned** in important respects, and from the **PADT annotations**.

Lexicon's Design

ElixirFM builds a **domain-specific embedded** language for the lexical data.

Lexicon's Design

ElixirFM builds a **domain-specific embedded** language for the lexical data.

- (a) representation of the linguistic data in an abstract and **extensible notation** that encodes both **orthography** and **phonology**, and whose interpretation is **customizable**

Lexicon's Design

ElixirFM builds a **domain-specific embedded** language for the lexical data.

- (a) representation of the linguistic data in an abstract and **extensible notation** that encodes both **orthography** and **phonology**, and whose interpretation is **customizable**
- (b) organization of the lexicon so that there is preferably **no duplication** of information and so that the lexicon can possibly be **divided** into separate units, as well as be **interlinked** with external **modules**

Lexicon's Design

ElixirFM builds a **domain-specific embedded** language for the lexical data.

- (a) representation of the linguistic data in an abstract and **extensible notation** that encodes both **orthography** and **phonology**, and whose interpretation is **customizable**
- (b) organization of the lexicon so that there is preferably **no duplication** of information and so that the lexicon can possibly be **divided** into separate units, as well as be **interlinked** with external **modules**
- (c) definition of such **format of the lexicon** so that editing and understanding the data is not inappropriately difficult, and using such data **markup** whose syntax is either **lightweight**, or can be edited/verified with some **automatic tools**, or both

|> "s l k" <| [

FiCL 'noun' ["wire", "thread"]
 'plural' HaFCAL,

FiCL |< Iy 'adj' ["wire", "by wire"],

lA >| FiCL |< Iy 'adj' ["wireless", "radio"]]

|> "s l k" <| [

FiCL 'noun' ["wire", "thread"]
 'plural' HaFCAL,

FiCL |< Iy 'adj' ["wire", "by wire"],

lA >| FiCL |< Iy 'adj' ["wireless", "radio"]]

"s l k" 'merge' al >| lA >| FiCL |< Iy |<< "u"

|> "s l k" <| [

FiCL 'noun' ["wire", "thread"]
 'plural' HaFCAL,

FiCL |< Iy 'adj' ["wire", "by wire"],

lA >| FiCL |< Iy 'adj' ["wireless", "radio"]]

"s l k" 'merge' al >| lA >| FiCL |< Iy |<< "u"

"al-lA-silkIyu"

|> "s l k" <| [

FiCL 'noun' ["wire", "thread"]
'plural' HaFCAL,

FiCL |< Iy 'adj' ["wire", "by wire"],

lA >| FiCL |< Iy 'adj' ["wireless", "radio"]]

"s l k" 'merge' al >| lA >| FiCL |< Iy |<< "u"

"al-lA-silkIyu"

الاسلكي اللّاسلكي al-lā-silkīyu اللّاسلكيُّ اللّاسلكيُّ

|> "s l k" <| [

FiCL 'noun' ["wire", "thread"]
'plural' HaFCAL,

FiCL |< Iy 'adj' ["wire", "by wire"],

lA >| FiCL |< Iy 'adj' ["wireless", "radio"]]

"s l k" 'merge' al >| lA >| FiCL |< Iy |<< "u"

"al-lA-silkIyu"

الاسلكي اللّاسلكي al-lā-silkīyu اللّاسلكيُّ اللّاسلكيُّ

Adjective Masculine Singular Nominative Definite

|> "s l k" <| [

FiCL 'noun' ["wire", "thread"]
'plural' HaFCAL,

FiCL |< Iy 'adj' ["wire", "by wire"],

lA >| FiCL |< Iy 'adj' ["wireless", "radio"]]

"s l k" 'merge' al >| lA >| FiCL |< Iy |<< "u"

"al-lA-silkIyu"

الاسلكي اللّاسلكي al-lā-silkīyu اللّاسلكي اللّاسلكي

Adjective Masculine Singular Nominative Definite

A-----MS1D

A-----MS1 [DC]

A-----MS1C

Morphology Disambiguation

Arabic is a language of **rich morphology**, both derivational and inflectional, with **highly ambiguous** orthography.

Morphology Disambiguation

Arabic is a language of **rich morphology**, both derivational and inflectional, with **highly ambiguous** orthography.

Boundaries of syntactic units, **tokens**, are obscure in writing—orthographic words, **strings**, consist of up to four **lexemes**.

Morphology Disambiguation

Arabic is a language of **rich morphology**, both derivational and inflectional, with **highly ambiguous** orthography.

Boundaries of syntactic units, **tokens**, are obscure in writing—orthographic words, **strings**, consist of up to four **lexemes**.

Disambiguation encompasses subproblems like **tokenization**

Morphology Disambiguation

Arabic is a language of **rich morphology**, both derivational and inflectional, with **highly ambiguous** orthography.

Boundaries of syntactic units, **tokens**, are obscure in writing—orthographic words, **strings**, consist of up to four **lexemes**.

Disambiguation encompasses subproblems like **tokenization**, **full morphological tagging** or its simplified '**part-of-speech**' versions

Morphology Disambiguation

Arabic is a language of **rich morphology**, both derivational and inflectional, with **highly ambiguous** orthography.

Boundaries of syntactic units, **tokens**, are obscure in writing—orthographic words, **strings**, consist of up to four **lexemes**.

Disambiguation encompasses subproblems like **tokenization**, **full morphological tagging** or its simplified '**part-of-speech**' versions, **lemmatization**

Morphology Disambiguation

Arabic is a language of **rich morphology**, both derivational and inflectional, with **highly ambiguous** orthography.

Boundaries of syntactic units, **tokens**, are obscure in writing—orthographic words, **strings**, consist of up to four **lexemes**.

Disambiguation encompasses subproblems like **tokenization**, **full morphological tagging** or its simplified '**part-of-speech**' versions, **lemmatization**, **diacritization** or restoration of the **structural components** of words

Morphology Disambiguation

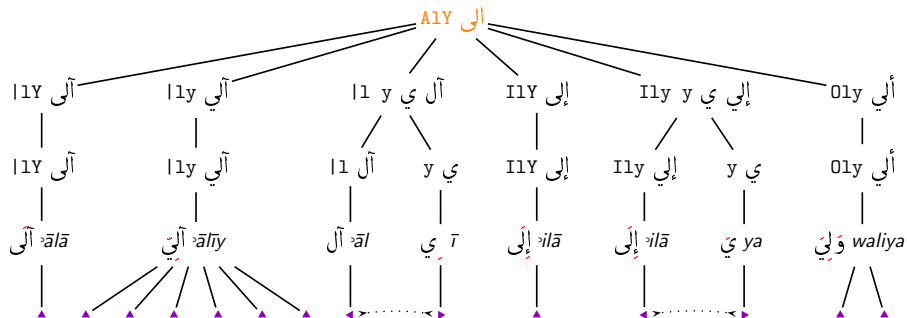
Arabic is a language of **rich morphology**, both derivational and inflectional, with **highly ambiguous** orthography.

Boundaries of syntactic units, **tokens**, are obscure in writing—orthographic words, **strings**, consist of up to four **lexemes**.

Disambiguation encompasses subproblems like **tokenization**, **full morphological tagging** or its simplified '**part-of-speech**' versions, **lemmatization**, **diacritization** or restoration of the **structural components** of words, **plus combinations** thereof.

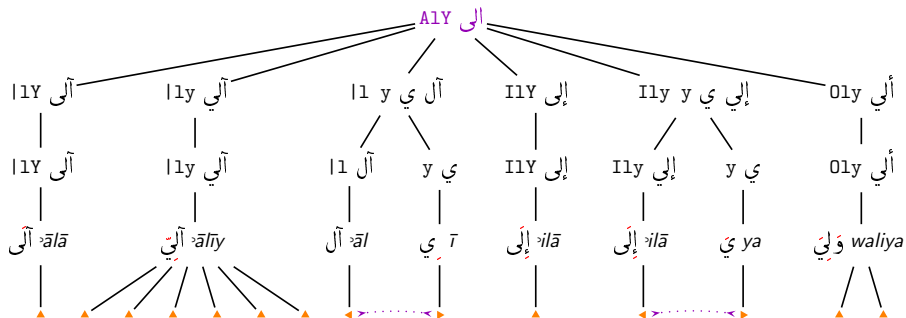
MorphoTrees

... organize the analyses into a hierarchy with the **string** as its **root**



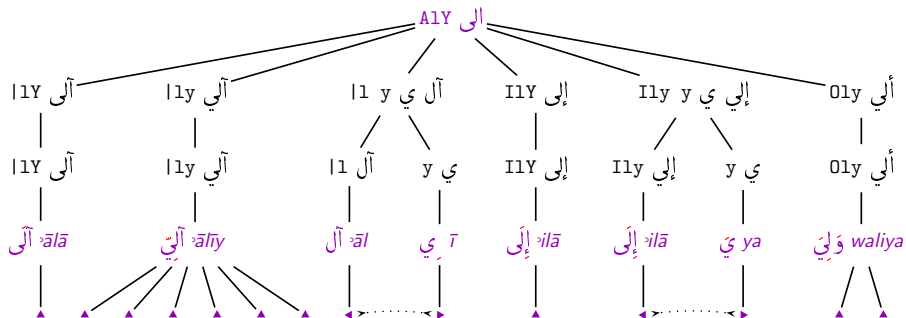
MorphoTrees

... organize the analyses into a hierarchy with the **string** as its **root** and the **full tokens** as the **leaves**



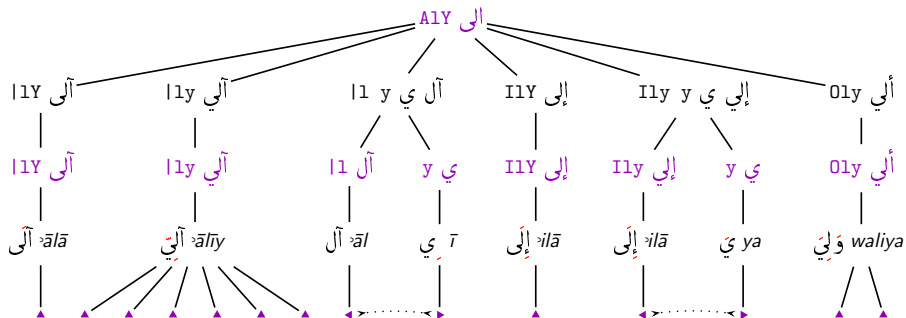
MorphoTrees

... organize the analyses into a hierarchy with the **string** as its **root** and the **full tokens** as the **leaves**, grouped by their **lemmas**



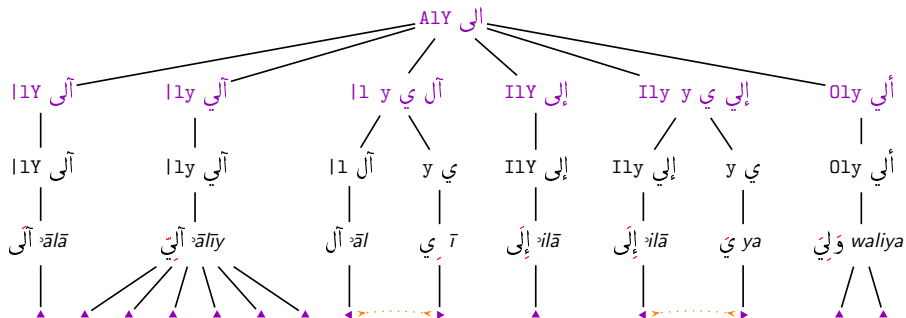
MorphoTrees

... organize the analyses into a hierarchy with the **string** as its **root** and the **full tokens** as the **leaves**, grouped by their **lemmas**, **canonical forms**

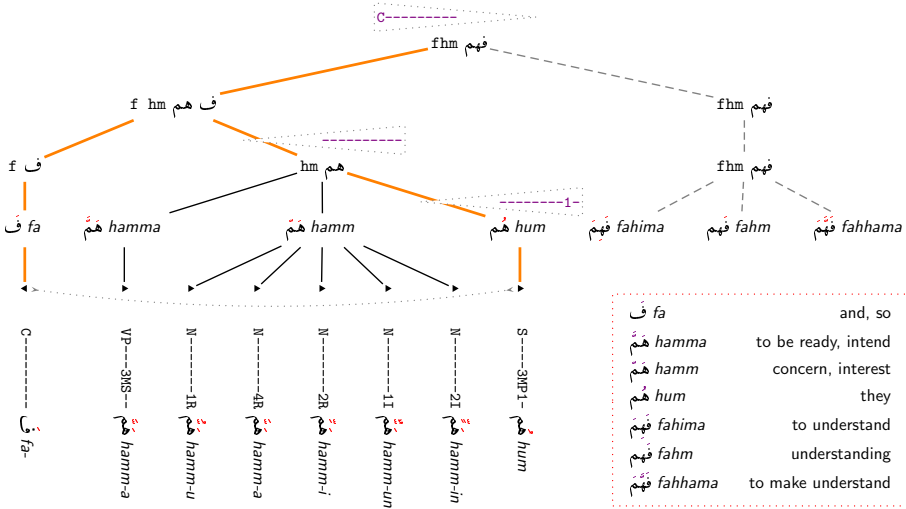


MorphoTrees

... organize the analyses into a hierarchy with the **string** as its **root** and the **full tokens** as the **leaves**, grouped by their **lemmas**, **canonical forms** and **partitionings** of the string into such forms:



Multi-Modal Annotation



Contributions

- (a) recognition of **functional versus illusory** morphological categories, definition of a **minimal but complete** system of inflectional parameters in Arabic

Contributions

- (a) recognition of **functional versus illusory** morphological categories, definition of a **minimal but complete** system of inflectional parameters in Arabic
- (b) **morphophonemic patterns** and their significance for the simplification of the model of morphological **alternations**

Contributions

- (a) recognition of **functional versus illusory** morphological categories, definition of a **minimal but complete** system of inflectional parameters in Arabic
- (b) **morphophonemic patterns** and their significance for the simplification of the model of morphological **alternations**
- (c) **inflectional invariant** and its consequence for the efficiency of morphological **recognition** in Arabic

Contributions

- (a) recognition of **functional versus illusory** morphological categories, definition of a **minimal but complete** system of inflectional parameters in Arabic
- (b) **morphophonemic patterns** and their significance for the simplification of the model of morphological **alternations**
- (c) **inflectional invariant** and its consequence for the efficiency of morphological **recognition** in Arabic
- (d) **intuitive notation** for the **structural components** of words

Contributions

- (a) recognition of **functional versus illusory** morphological categories, definition of a **minimal but complete** system of inflectional parameters in Arabic
- (b) **morphophonemic patterns** and their significance for the simplification of the model of morphological **alternations**
- (c) **inflectional invariant** and its consequence for the efficiency of morphological **recognition** in Arabic
- (d) **intuitive notation** for the **structural components** of words
- (e) conversion of the **Buckwalter lexicon** into a functional format resembling **printed dictionaries**

Contributions

- (f) ElixirFM as a **general-purpose** model of morphological **inflection and derivation** in Arabic, implemented with **high-level** declarative programming

Contributions

- (f) ElixirFM as a **general-purpose** model of morphological **inflection and derivation** in Arabic, implemented with **high-level** declarative programming
- (g) abstraction from one particular **orthography** affecting the clarity of the model and extending its applicability to **other written representations** of the language

Contributions

- (f) ElixirFM as a **general-purpose** model of morphological **inflection and derivation** in Arabic, implemented with **high-level** declarative programming
- (g) abstraction from one particular **orthography** affecting the clarity of the model and extending its applicability to **other written representations** of the language
- (h) **MorphoTrees** as a **hierarchization** of the process of morphological disambiguation

Contributions

- (f) ElixirFM as a **general-purpose** model of morphological **inflection and derivation** in Arabic, implemented with **high-level** declarative programming
- (g) abstraction from one particular **orthography** affecting the clarity of the model and extending its applicability to **other written representations** of the language
- (h) **MorphoTrees** as a **hierarchization** of the process of morphological disambiguation
- (i) **expandable** morphological positional tags, **restrictions** on features, their **inheritance**

Contributions

- (f) ElixirFM as a **general-purpose** model of morphological **inflection and derivation** in Arabic, implemented with **high-level** declarative programming
- (g) abstraction from one particular **orthography** affecting the clarity of the model and extending its applicability to **other written representations** of the language
- (h) **MorphoTrees** as a **hierarchization** of the process of morphological disambiguation
- (i) **expandable** morphological positional tags, **restrictions** on features, their **inheritance**
- (j) **open-source implementations** of ElixirFM, **Encode Arabic**, MorphoTrees, and **extensions for Arab \TeX**

ElixirFM and its lexicons are **open-source software** licensed under GNU GPL

<http://sourceforge.net/projects/elixir-fm/>

Evaluation

problem inter-annotator agreement MorphoTrees vs. MorphoLists

words 1215

tokens 1400

chance 397

disagreement MorphoTrees 5.3% MorphoLists 9.3%

Annotator A 2.8% Annotator B 5.9%

Coverage

Buckwalter errors in explicit 'paradigmatic' listings, e.g. ʾabāna → **tubinū* تبنوا

passives and imperatives **not generated** by rules either

Coverage

Buckwalter errors in explicit 'paradigmatic' listings, e.g. ʾabāna → **tubinū* تبنوا

passives and imperatives **not generated** by rules either

ElixirFM **roots and patterns** explicit, but perhaps excessive or incorrectly inferred

jussive and energetic moods **supported fully**

deverbal nouns and participles can often **be derived**