

Tips and Tricks of the Prague Arabic Dependency Treebank

Otakar Smrž

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

The Challenge of Arabic for NLP/MT

The British Computer Society, London, 23 October 2006

Prague Arabic Dependency Treebank

PADT is a project of linguistic annotation of **Modern Written Arabic** based on the theory of **Functional Generative Description**.

Prague Arabic Dependency Treebank

PADT is a project of linguistic annotation of **Modern Written Arabic** based on the theory of **Functional Generative Description**.

PADT 1.0 was published in 2004 by the Linguistic Data Consortium, and has been used by tens of academic and commercial institutions.

Prague Arabic Dependency Treebank

PADT is a project of linguistic annotation of **Modern Written Arabic** based on the theory of **Functional Generative Description**.

PADT 1.0 was published in 2004 by the Linguistic Data Consortium, and has been used by tens of academic and commercial institutions.

In PADT, which now consists of the **morphological** and the **analytical** levels of description of Arabic, the annotation of **tectogrammatics** and **information structure** is being established.

Outline

1 Introduction

Outline

- 1 Introduction
- 2 Text Processing
 - Buckwalter Transliteration
 - Notation of ArabT_EX
 - Encode Arabic

Outline

- 1 Introduction
- 2 Text Processing
 - Buckwalter Transliteration
 - Notation of ArabT_EX
 - Encode Arabic
- 3 Morphology
 - MorphoTrees
 - ElixirFM

Outline

- 1 Introduction
- 2 Text Processing
 - Buckwalter Transliteration
 - Notation of ArabT_EX
 - Encode Arabic
- 3 Morphology
 - MorphoTrees
 - ElixirFM
- 4 Syntax and Beyond
 - Analytical Syntax
 - Tectogrammatics

Outline

- 1 Introduction
- 2 Text Processing
 - Buckwalter Transliteration
 - Notation of ArabT_EX
 - Encode Arabic
- 3 Morphology
 - MorphoTrees
 - ElixirFM
- 4 Syntax and Beyond
 - Analytical Syntax
 - Tectogrammatics
- 5 TrEd Live

Outline

- 1 Introduction
- 2 Text Processing
 - Buckwalter Transliteration
 - Notation of ArabT_EX
 - Encode Arabic
- 3 Morphology
 - MorphoTrees
 - ElixirFM
- 4 Syntax and Beyond
 - Analytical Syntax
 - Tectogrammatics
- 5 TrEd Live
- 6 References

Outline

- 1 Introduction
- 2 Text Processing
 - Buckwalter Transliteration
 - Notation of ArabT_EX
 - Encode Arabic
- 3 Morphology
 - MorphoTrees
 - ElixirFM
- 4 Syntax and Beyond
 - Analytical Syntax
 - Tectogrammatcs
- 5 TrEd Live
- 6 References

Buckwalter Transliteration

يُولَدُ جَمِيعُ النَّاسِ أَحْرَارًا مُتَسَاوِينَ فِي الْكِرَامَةِ وَالْحُقُوقِ. وَقَدْ وَهَبُوا عَقْلًا وَضَمِيرًا وَعَلَيْهِمْ
أَنْ يُعَامَلَ بَعْضُهُمْ بَعْضًا بِرُوحِ الْإِحَاءِ.

yuwladu jamiyEu {ln~aAsi OaHoraArFA mutasaAwiyna fiy
{lokaraAmapi wa {loHuquwqi. waqado wuhibuWA EaqlAF
waDamiyrFA waEalayohimo Oano yuEaAmila baEoDuhumo baEoDFA
biruwHi {loIixaA'i.

Buckwalter Transliteration

يولد جميع الناس أحرارا متساوين في الكرامة والحقوق. وقد وهبوا عقلا وضميرا وعليهم أن يعامل بعضهم بعضا بروح الإخاء.

ywld jmyE AlnAs OHrArA mtsAwyn fy AlkrAmp wAlHqwq. wqd
whbWA EqLA wDmyrA wElyhm On yEAml bEDhm bEDA brwH AlIxA'.

Buckwalter Transliteration

يُولَدُ جَمِيعُ النَّاسِ أحرَارًا مُتساوِينَ فِي الكَرَامَةِ وَالْحُقُوقِ. وَقَدْ وَهَبُوا عَقْلاً وَضَمِيرًا وَعَلَيْهِمْ
أَنْ يُعَامَلَ بَعْضُهُمْ بَعْضًا بِرُوحِ الإِخَاءِ.

yuwladu jamiyEu {ln~aAsi OaHORAarFA mutasaAwiyna fiy
{lokaraAmapI wa {loHuquwqi. waqado wuhibuWA EaQoLAF
waDamIyrFA waEalayohimo Oano yuEaAmila baEoDuhumo baEoDFA
biruwHi {loIixaA'i.

يولد جميع الناس أحرارا متساوين في الكرامة والحقوق. وقد وهبوا عقلا وضميرا وعليهم
أن يعامل بعضهم بعضا بروح الإخاء.

ywld jmyE AlnAs OHrArA mtsAwyn fy AlkrAmp wAlHqwq. wqd
whbWA EqLA wDmyrA wElyhm On yEAml bEDhm bEDA brwH AlIxA'.

Notation of Arab_TE_X

يُولَدُ جَمِيعُ النَّاسِ أَحْرَارًا مُتَسَاوِينَ فِي الْكِرَامَةِ وَالْحُقُوقِ. وَقَدْ وَهَبُوا عَقْلًا وَضَمِيرًا وَعَلَيْهِمْ
أَنْ يُعَامَلَ بَعْضُهُمْ بَعْضًا بِرُوحِ الْإِخَاءِ.

```
\cap yUladu ^gamI'u an-nAsi 'a.hrAraN mutasAwIna fI
al-karAmaTi wa-al-.huqUqi.
```

```
\cap wa-qad wuhibUA 'aqlaN wa-.damIraN wa-'alayhim 'an
yu'Amila ba'.duhum ba'.daN bi-rU.hi al-'i_hA'i.
```

Notation of ArabT_EX

يولد جميع الناس أحرارا متساوين في الكرامة والحقوق. وقد وهبوا عقلا وضميرا وعليهم
أن يعامل بعضهم بعضا بروح الإخاء.

```
\cap yUladu ^gamI'u an-nAsi 'a.hrAraN mutasAwIna fI
al-karAmaTi wa-al-.huqUqi.
```

```
\cap wa-qad wuhibUA 'aqlaN wa-.damIraN wa-'alayhim 'an
yu'Amila ba'.duhum ba'.daN bi-rU.hi al-'i_hA'i.
```


Notation of ArabT_EX

Yūladu ḡamīʿu ʿn-nāsi ʿaḥrāran mutasāwīna fī ʿl-karāmati wa-ʿl-ḥuqūqi. Wa-qad wuhibū ʿaqlan wa-ḍamīran wa-ʿalayhim ʿan yuʿāmila baḍduhum baḍdan bi-rūḥi ʿl-ʿiḥāʿi.

```
\cap yUladu ^gamI'u an-nAsi 'a.hrAraN mutasAwIna fI
al-karAmaTi wa-al-.huqUqi.
```

```
\cap wa-qad wuhibUA 'aqlaN wa-.damIraN wa-'alayhim 'an
yu'Amila ba'.duhum ba'.daN bi-rU.hi al-'i_hA'i.
```

Notation of ArabT_EX

يُولَدُ جَمِيعُ النَّاسِ أَحْرَارًا مُتَسَاوِينَ فِي الْكَرَامَةِ وَالْحُقُوقِ. وَقَدْ وَهَبُوا عَقْلًا وَضَمِيرًا وَعَلَيْهِمْ أَنْ يُعَامَلَ بَعْضُهُمْ بَعْضًا بِرُوحِ الْإِحَاءِ.

يولد جميع الناس أحرارا متساوين في الكرامة والحقوق. وقد وهبوا عقلا وضميرا وعليهم أن يعامل بعضهم بعضا بروح الإخاء.

Yūladu ġamī'u 'n-nāsi 'aħrāran mutasāwīna fī 'l-karāmati wa-'l-ħuqūqi. Wa-qad wuhibū 'aqlan wa-ḍamīran wa-ʿalayhim 'an yuʿāmila baḍduhum baḍdan bi-rūḥi 'l-iḥā'i.

```
\cap yUladu ^gamI'u an-nAsi 'a.hrAraN mutasAwIna fI
al-karAmaTi wa-al-.huqUqi.
```

```
\cap wa-qad wuhibUA 'aqlaN wa-.damIraN wa-'alayhim 'an
yu'Amila ba'.duhum ba'.daN bi-rU.hi al-'i_hA'i.
```

Encode Arabic

biruwHi {loIixaA'i ←  ← bi-rU.hi al-'i_hA'i

Implemented in **Perl** and available on CPAN as **Encode-Arabic**:

```
$encoded = encode "buckwalter", decode "arabtex", $decoded
$encoded = encode("buckwalter", decode("arabtex", $decoded))
```

Implemented in **Haskell** and available along with **ElixirFM**:

```
encoded = encode Buckwalter $ decode ArabTeX decoded
encoded = encode Buckwalter (decode ArabTeX decoded)
encoded = (encode Buckwalter . decode ArabTeX) decoded
```

```
[cmd] decode ArabTeX < decode.d | encode Buckwalter > encode.d
```

Encode Arabic

biruwHi {loIixaA'i ← بِرُوحِ الْإِخَاءِ ← bi-rU.hi al-'i_hA'i

Implemented in **Perl** and available on CPAN as **Encode-Arabic**:

```
$encoded = encode "buckwalter", decode "arabtex", $decoded
$encoded = encode("buckwalter", decode("arabtex", $decoded))
```

Implemented in **Haskell** and available along with **ElixirFM**:

```
encoded = encode Buckwalter $ decode ArabTeX decoded
encoded = encode Buckwalter (decode ArabTeX decoded)
encoded = (encode Buckwalter . decode ArabTeX) decoded
```

```
[cmd] decode ArabTeX < decode.d | encode Buckwalter > encode.d
```

Encode Arabic

biruwHi {loIixaA'i ←  ← bi-rU.hi al-'i_hA'i

Implemented in **Perl** and available on CPAN as **Encode-Arabic**:

```
$encoded = encode "buckwalter", decode "arabtex", $decoded
$encoded = encode("buckwalter", decode("arabtex", $decoded))
```

Implemented in **Haskell** and available along with **ElixirFM**:

```
encoded = encode Buckwalter $ decode ArabTeX decoded
encoded = encode Buckwalter (decode ArabTeX decoded)
encoded = (encode Buckwalter . decode ArabTeX) decoded
```

```
[cmd] decode ArabTeX < decode.d | encode Buckwalter > encode.d
```

Encode Arabic

biruwHi {loIixaA'i ←  ← bi-rU.hi al-'i_hA'i

Implemented in **Perl** and available on CPAN as **Encode-Arabic**:

```
$encoded = encode "buckwalter", decode "arabtex", $decoded
$encoded = encode("buckwalter", decode("arabtex", $decoded))
```

Implemented in **Haskell** and available along with **ElixirFM**:

```
encoded = encode Buckwalter $ decode ArabTeX decoded
encoded = encode Buckwalter (decode ArabTeX decoded)
encoded = (encode Buckwalter . decode ArabTeX) decoded
```

```
[cmd] decode ArabTeX < decode.d | encode Buckwalter > encode.d
```

Encode Arabic

biruwHi {loIixaA'i ← ← bi-rU.hi al-'i_hA'i

Implemented in **Perl** and available on CPAN as **Encode-Arabic**:

```
$encoded = encode "buckwalter", decode "arabtex", $decoded
$encoded = encode("buckwalter", decode("arabtex", $decoded))
```

Implemented in **Haskell** and available along with **ElixirFM**:

```
encoded = encode Buckwalter $ decode ArabTeX decoded
encoded = encode Buckwalter (decode ArabTeX decoded)
encoded = (encode Buckwalter . decode ArabTeX) decoded
```

`[cmd] decode ArabTeX < decode.d | encode Buckwalter > encode.d`

Outline

- 1 Introduction
- 2 Text Processing
 - Buckwalter Transliteration
 - Notation of ArabT_EX
 - Encode Arabic
- 3 Morphology**
 - MorphoTrees
 - ElixirFM
- 4 Syntax and Beyond
 - Analytical Syntax
 - Tectogrammatics
- 5 TrEd Live
- 6 References

He will notify them about that through SMS messages, the Internet, and other means. سيخبرهم بذلك عن طريق الرسائل القصيرة والإنترنت وغيرها.

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

He will notify them about that through SMS messages, the Internet, and other means. سَيُخَبِّرُهُمْ بِذَلِكَ عَنِ طَرِيقِ الرَّسَائِلِ الْقَصِيرَةِ وَالْإِنْتَرْنِتِ وَغَيْرِهَا.

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

He will notify them about that through SMS messages, the Internet, and other means. سَيُخَبِّرُهُمْ بِذَلِكَ عَنِ طَرِيقِ الرَّسَائِلِ الْقَصِيرَةِ وَالْإِنْتَرِنِتِ وَغَيْرِهَا.

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

He will notify them about that through SMS messages, the Internet, and other means. سَيُخَبِّرُهُمْ بِذَلِكَ عَنْ طَرِيقِ الرَّسَائِلِ الْقَصِيرَةِ وَالْإِنْتَرِنِتِ وَغَيْرِهَا.

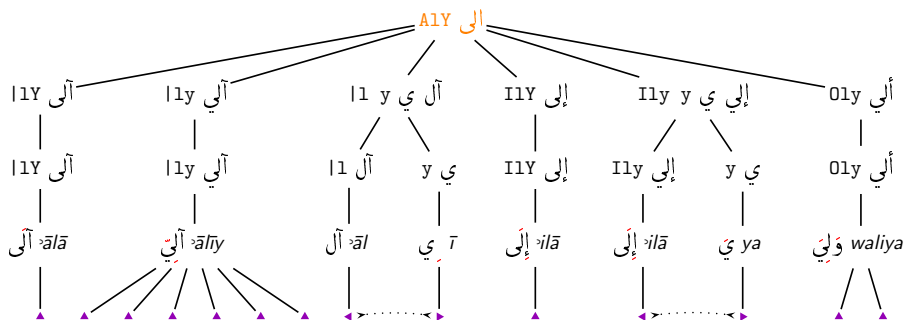
String	Token	Token Tag	Buckwalter's M-Tags	Token Form	Token Gloss
	F-----	FUT		<i>sa-</i>	will
سيخبرهم	VIIA-3MS--	IV3MS+IV+IVSUFF_MOOD:I		<i>yu-ḥbir-u</i>	he-notify
	S----3MP4-	IVSUFF_DO:3MS		<i>-hum</i>	them
بذلك	P-----	PREP		<i>bi-</i>	about/by
	SD----MS--	DEM_PRON_MS		<i>dālika</i>	that
عن	P-----	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-rasā'il-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

MorphoTrees

Suppose you can list **morphological analyses** for a given **input string** ...

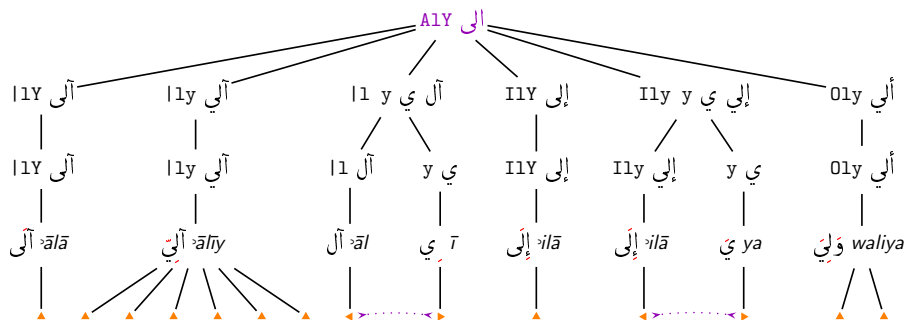
MorphoTrees

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



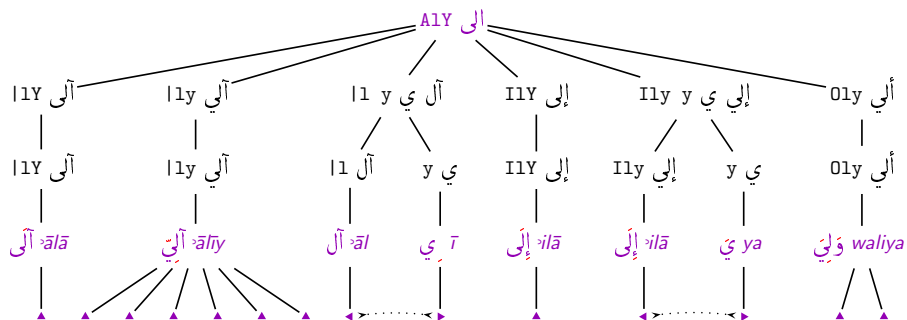
MorphoTrees

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



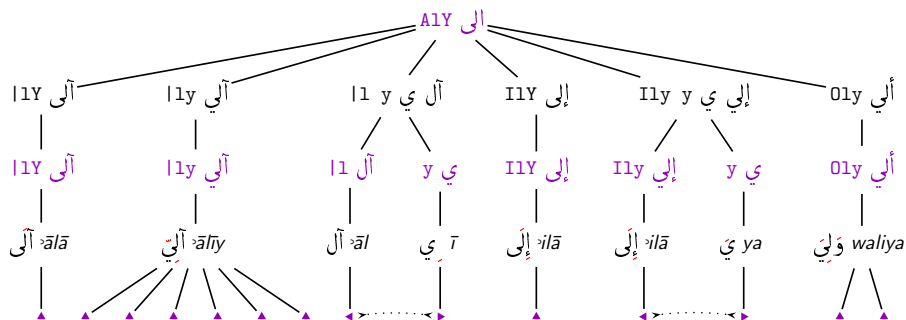
MorphoTrees

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



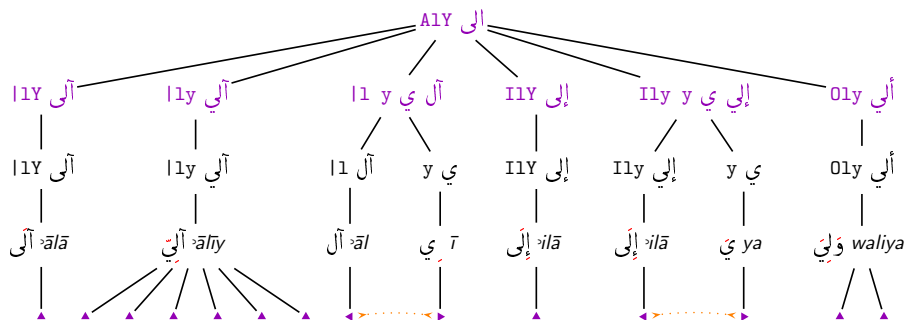
MorphoTrees

... 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**

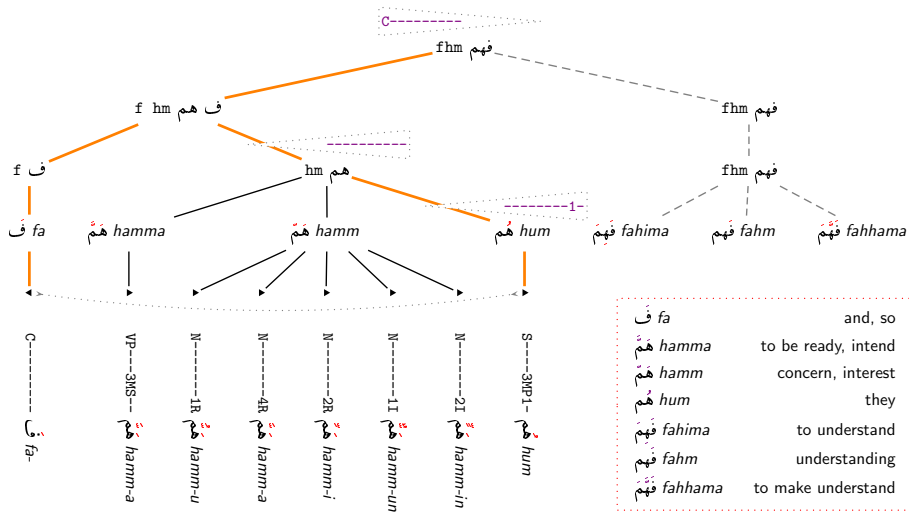


MorphoTrees

... 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:



MorphoTrees



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.

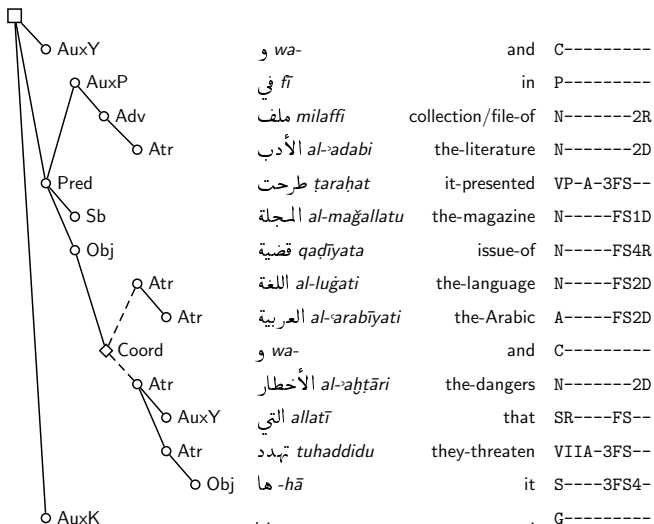
Morphology is **modeled** in terms of **paradigms**, grammatical **categories**, **lexemes** and word **classes**. The **computation** of analysis or generation is conceptually **distinguished** from the **general-purpose** linguistic **model**.

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

Outline

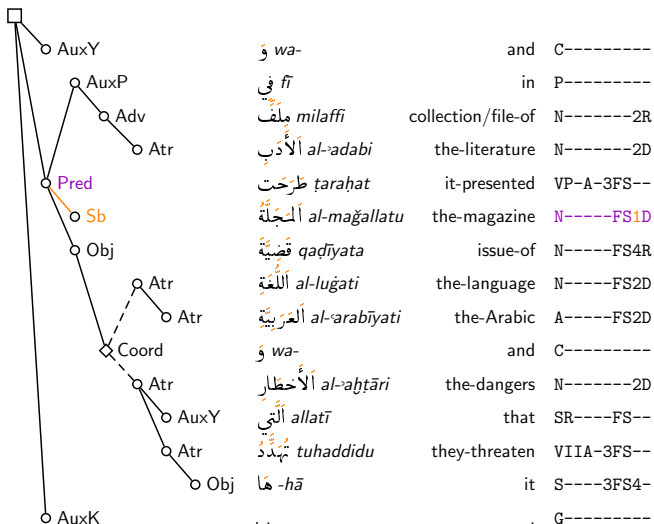
- 1 Introduction
- 2 Text Processing
 - Buckwalter Transliteration
 - Notation of ArabT_EX
 - Encode Arabic
- 3 Morphology
 - MorphoTrees
 - ElixirFM
- 4 Syntax and Beyond
 - Analytical Syntax
 - Tectogrammatics
- 5 TrEd Live
- 6 References

In the section on literature, the magazine presented the issue of the Arabic language and the dangers that threaten it. . . . وفي ملف الأدب طرحت



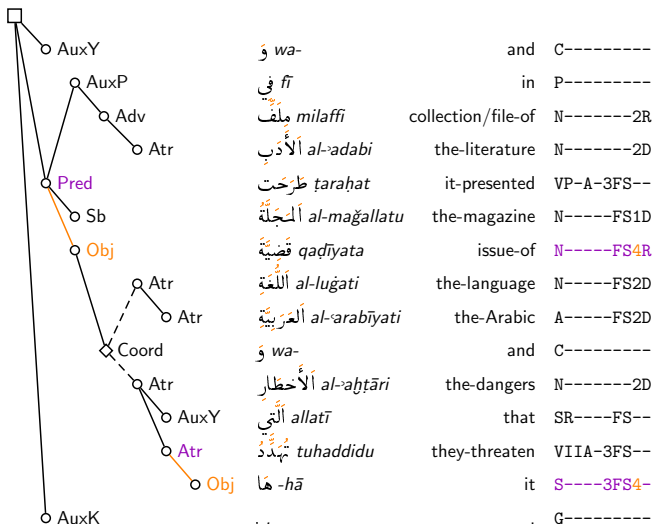
In the section on literature, the magazine presented the issue of the Arabic language and the dangers that threaten it. . . .

وَفِي مِلْفِ الْأَدَبِ طَرَحَتْ . . .



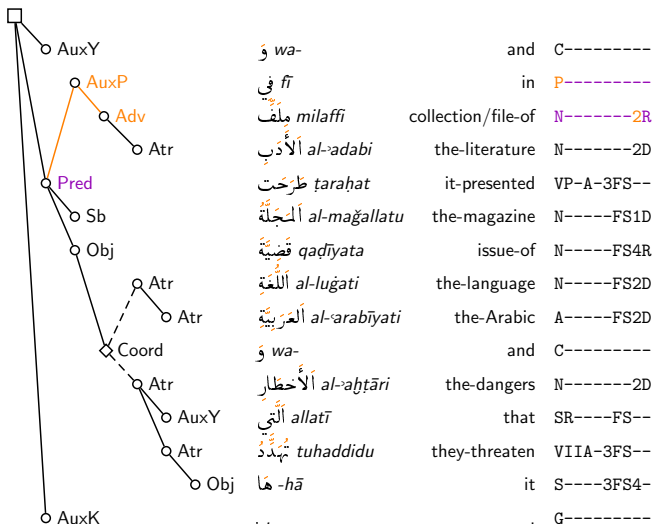
In the section on literature, the magazine presented the issue of the Arabic language and the dangers that threaten it. . . .

وَفِي مِلْفِ الْأَدَبِ طَرَحَتْ . . .

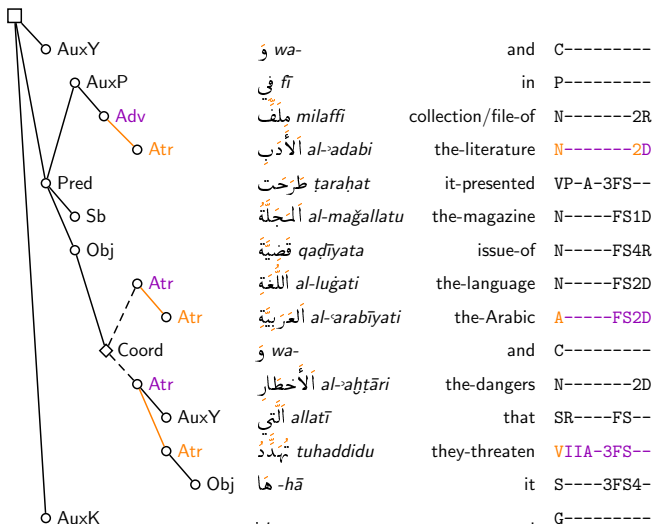


In the section on literature, the magazine presented the issue of the Arabic language and the dangers that threaten it. . . .

وَفِي مِلْفِ الْأَدَبِ طَرَحَتْ . . .

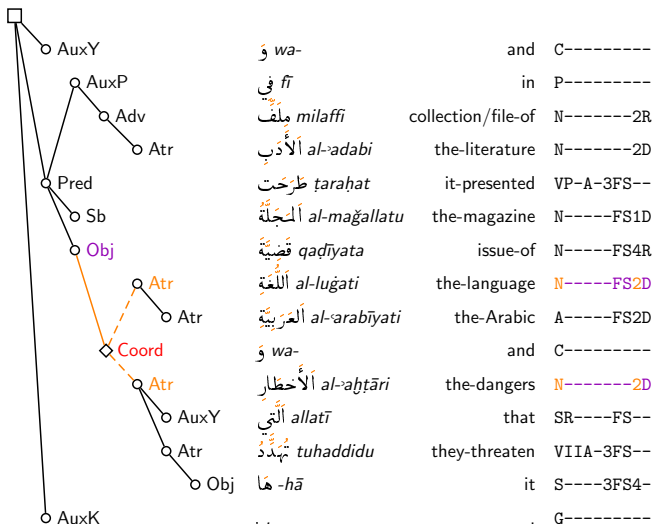


In the section on literature, the magazine presented the issue of the Arabic language and the dangers that threaten it. . . . وَفِي مِلْفِ الْأَدَبِ طَرَحَتْ . . .



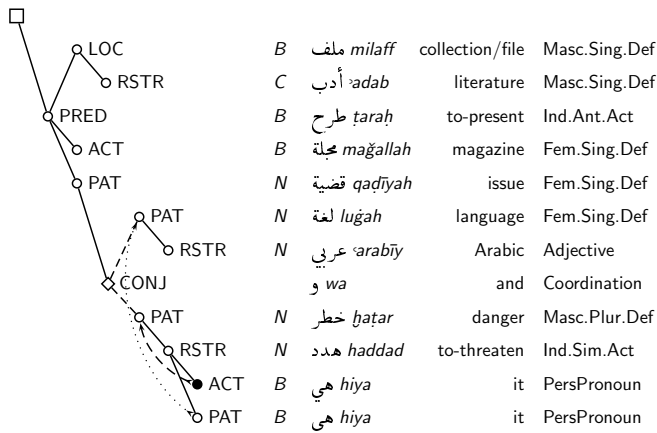
In the section on literature, the magazine presented the issue of the Arabic language and the dangers that threaten it. . . .

وَفِي مِلْفِ الْأَدَبِ طَرَحَتْ . . .



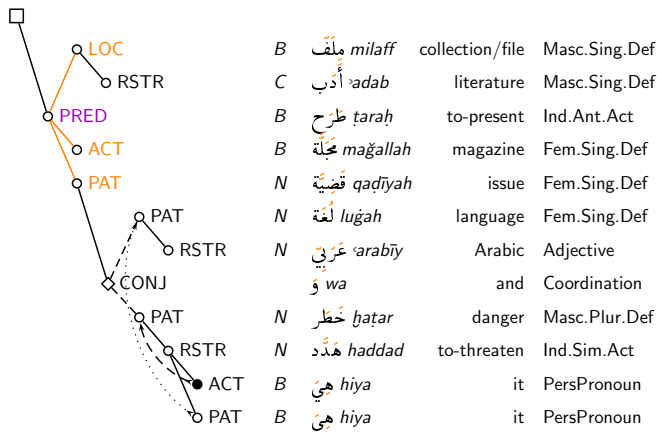
Tectogrammatics

Description of **linguistic meaning** in its **semantic** and **pragmatic** aspects.



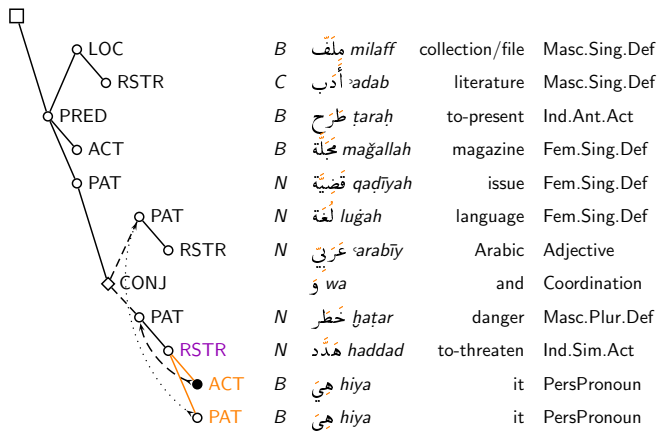
Tectogrammatics

Description of **linguistic meaning** in its **semantic** and **pragmatic** aspects.



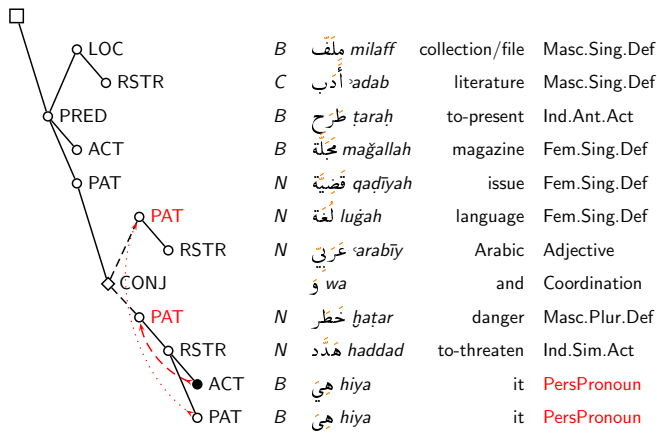
Tectogrammatics

Description of **linguistic meaning** in its **semantic** and **pragmatic** aspects.



Tectogrammatics

Description of **linguistic meaning** in its **semantic** and **pragmatic** aspects.



Outline

- 1 Introduction
- 2 Text Processing
 - Buckwalter Transliteration
 - Notation of ArabT_EX
 - Encode Arabic
- 3 Morphology
 - MorphoTrees
 - ElixirFM
- 4 Syntax and Beyond
 - Analytical Syntax
 - Tectogrammatics
- 5 TrEd Live**
- 6 References

Outline

- 1 Introduction
- 2 Text Processing
 - Buckwalter Transliteration
 - Notation of ArabT_EX
 - Encode Arabic
- 3 Morphology
 - MorphoTrees
 - ElixirFM
- 4 Syntax and Beyond
 - Analytical Syntax
 - Tectogrammatics
- 5 TrEd Live
- 6 **References**

- Buckwalter, Tim. [Buckwalter Arabic Morphological Analyzer 1.0](#). LDC catalog number LDC2002L49, ISBN 1-58563-257-0. 2002
- Forsberg, Markus and Aarne Ranta. [Functional Morphology](#). Proceedings of ICFP 2004, pages 213–223. ACM Press. 2004
- Lagally, Klaus. [ArabTeX: Typesetting Arabic and Hebrew, User Manual Version 4.00](#). Technical Report 2004/03, Fakultät Informatik, Universität Stuttgart. 2004
- Sgall, Petr and Eva Hajičová and Jarmila Panevová. [The Meaning of the Sentence in Its Semantic and Pragmatic Aspects](#). Academia, Prague. 1986
- Smrž, Otakar and Petr Pajas. [MorphoTrees of Arabic and Their Annotation in the TrEd Environment](#). Proceedings of the NEMLAR Conference 2004, pages 38–41. 2004

[PADT++](http://ufal.mff.cuni.cz/padt/online/) <http://ufal.mff.cuni.cz/padt/online/>