Prague Treebanking for Everyone

Prague Arabic Dependency Treebank

Otakar Smrž

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

Vilem Mathesius Lecture Series 21

Prague, November 29, 2006

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

Outline

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—orthographical 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.

He will notify them about that through SMS messages, the Internet, and other means. سَيُخبِرُهُم بِذٰلِكَ عَن طَرِيقِ الرَّسَائِلِ القَصِيرَةِ وَالإِنتِرنِتِ وَغَيرِهَا.

String	Token	Token Tag	Buckwalter's M-Tags	Token Form	Token Gloss
سيخبرهم		F---------	FUT	*sa-*	will
		VIIA-3MS--	IV3MS+IV+IVSUFF_MOOD:I	*yu-ḫbir-u*	he-notify
		S----3MP4-	IVSUFF_DO:3MP	*-hum*	them
بذلك		P---------	PREP	*bi-*	about/by
		SD----MS--	DEM_PRON_MS	*ḏālika*	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-rasāʾil-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	*ġayr-i*	other/not-of
		S----3FS2-	POSS_PRON_3FS	*-hā*	them

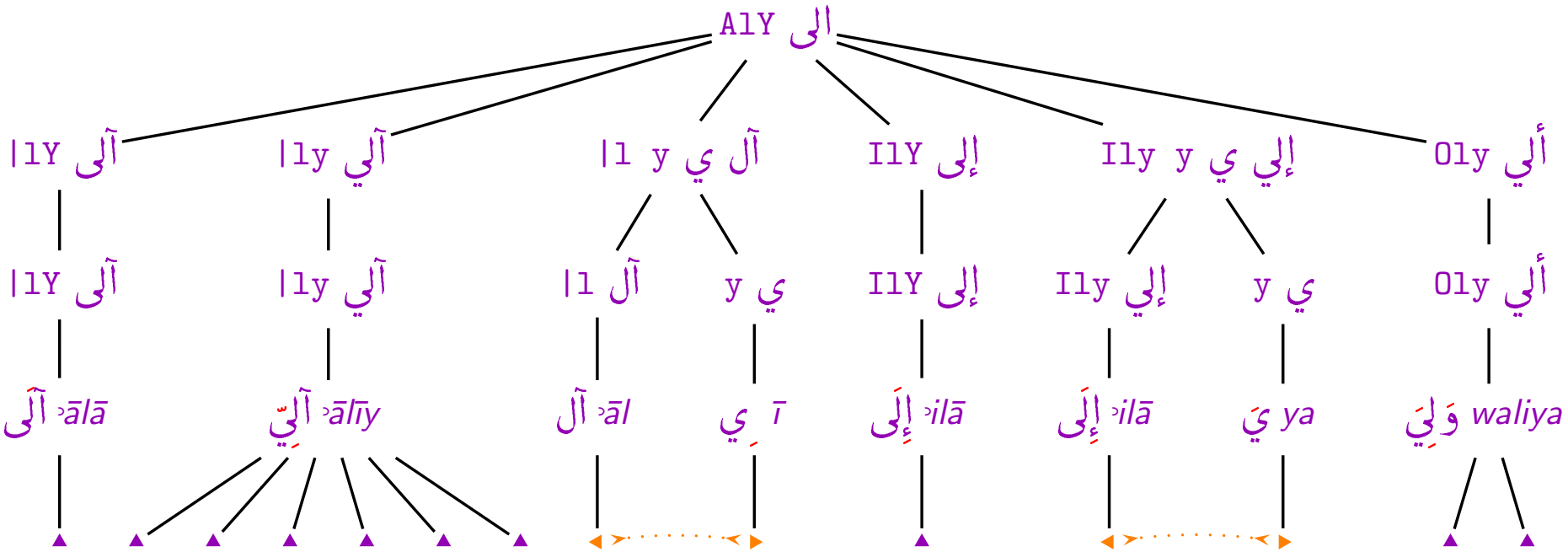
MorphoTrees

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

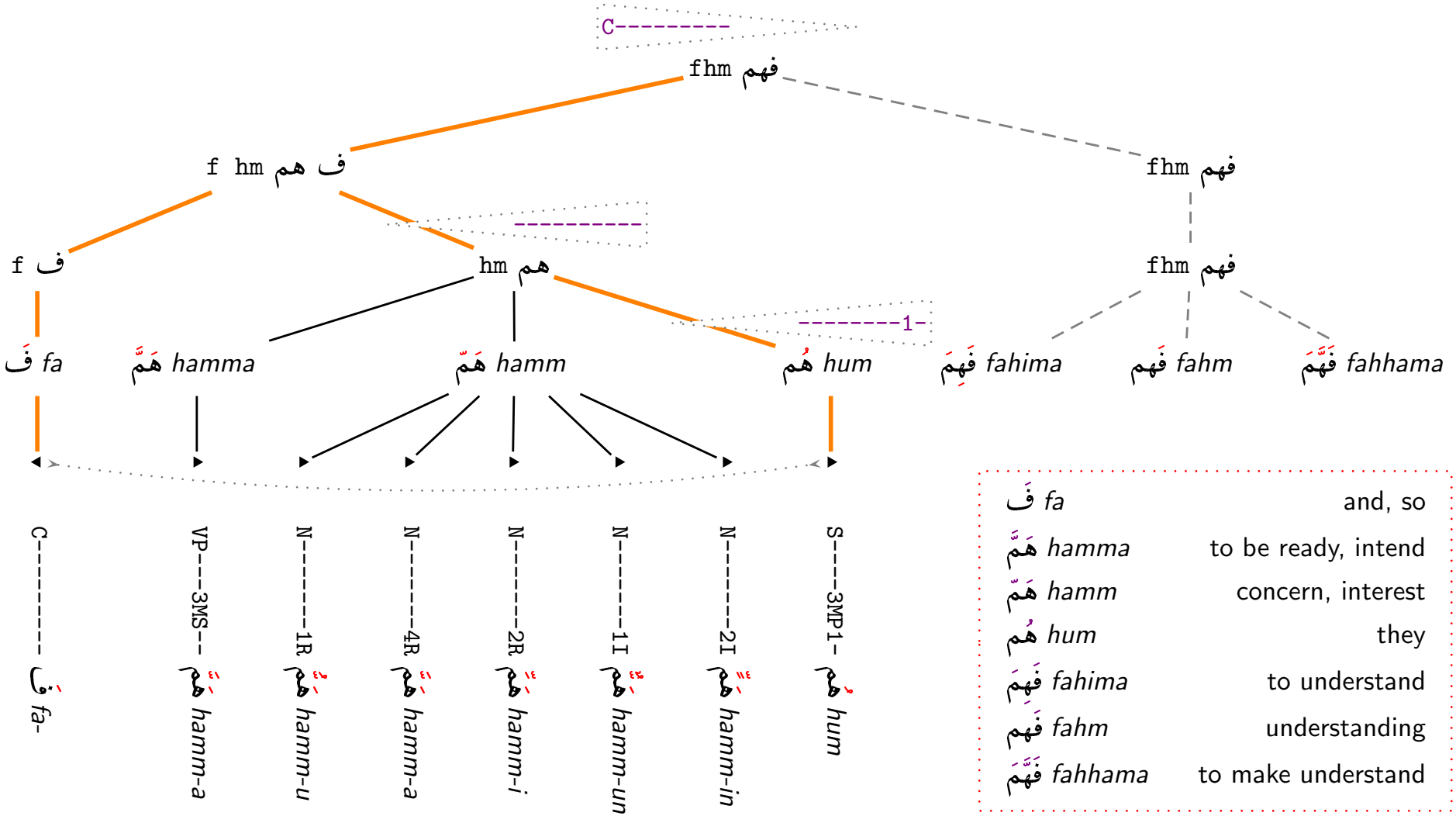
Morphs	Form	Token Tag	Lemma	Glosses per Morph
|laY+(null)	ʾālā	VP-A-3MS--	ʾālā	promise/take an oath + he/it
|liy˜	ʾālīy	A---------	ʾālīy	mechanical/automatic
|liy˜+u	ʾālīy-u	A-------1R	ʾālīy	mechanical . . . + [def.nom.]
|liy˜+i	ʾālīy-i	A-------2R	ʾālīy	mechanical . . . + [def.gen.]
|liy˜+a	ʾālīy-a	A-------4R	ʾālīy	mechanical . . . + [def.acc.]
|liy˜+N	ʾālīy-un	A-------1I	ʾālīy	mechanical . . . + [indef.nom.]
|liy˜+K	ʾālīy-in	A-------2I	ʾālīy	mechanical . . . + [indef.gen.]
|l +	ʾāl	N--------R	ʾāl	family/clan +
+ iy	-ī	S----1-S2-	ī	+ my
IilaY	ʾilā	P---------	ʾilā	to/towards
Iilay +	ʾilay	P---------	ʾilā	to/towards +
+ ya	-ya	S----1-S2-	ya	+ me
Oa+liy+(null)	ʾa-lī	VIIA-1-S--	waliya	I + follow/come after + [ind.]
Oa+liy+a	ʾa-liy-a	VISA-1-S--	waliya	I + follow/come after + [sub.]

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



فَ *fa*	and, so
هَمَّ *hamma*	to be ready, intend
هَمّ *hamm*	concern, interest
هُم *hum*	they
فَهِمَ *fahima*	to understand
فَهم *fahm*	understanding
فَهَّمَ *fahhama*	to make understand

ElixirFM

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

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

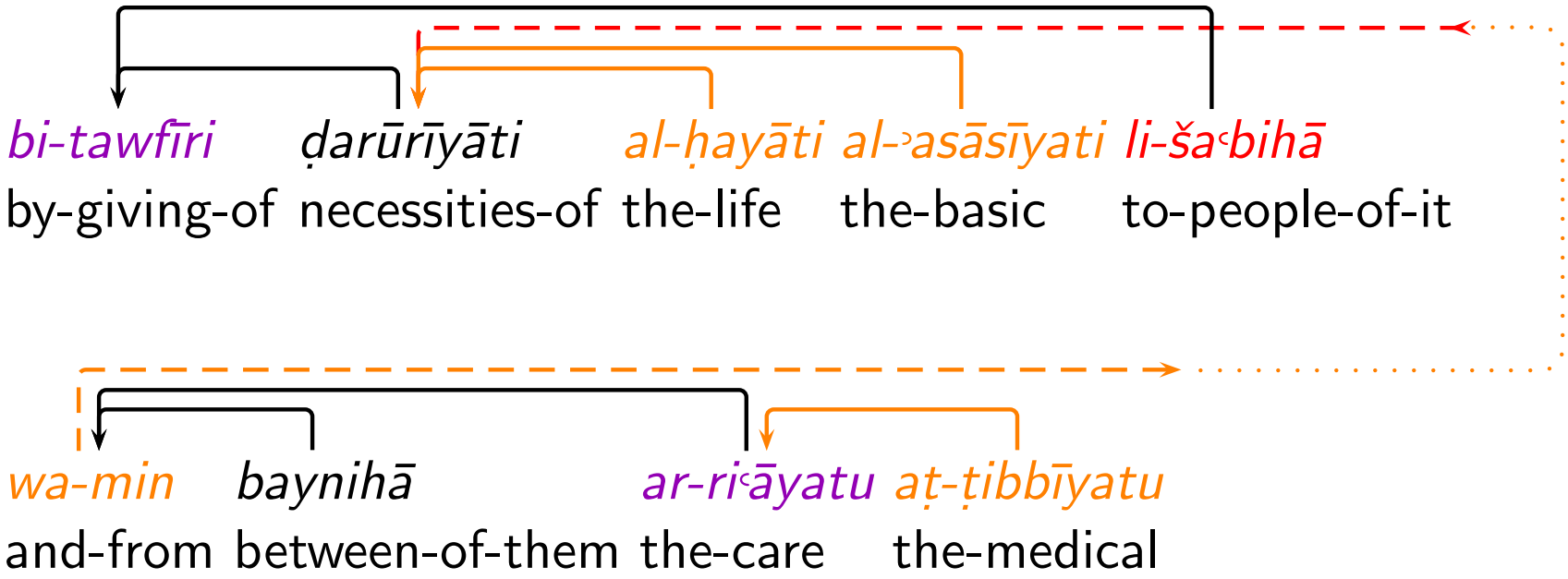
Morphology is modeled in terms of paradigms, grammatical categories, lexemes and word classes. The computation of analysis or generation is conceptually distinguished from the general-purpose linguistic model.

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

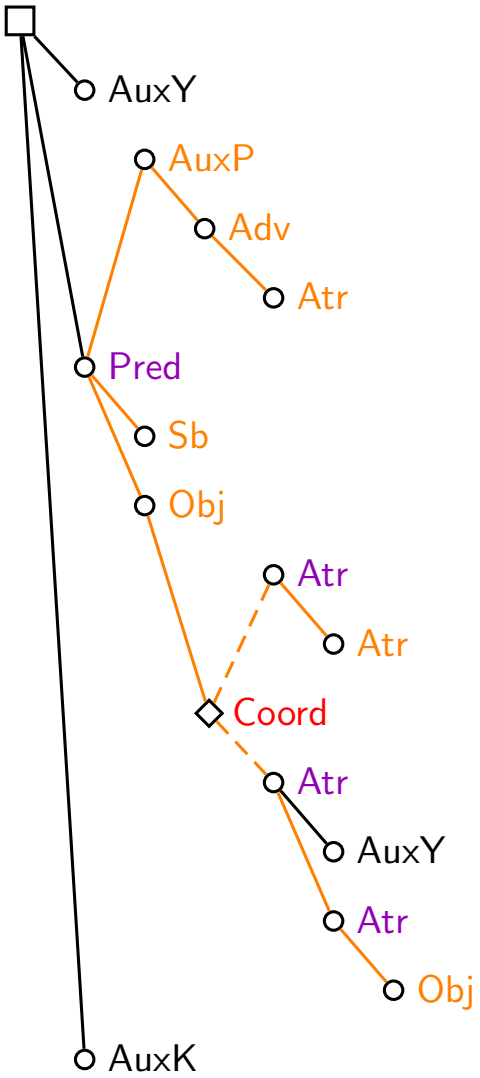
Outline

Dependency vs. Linearity

. . . by providing the basic necessities of life to its people, including medical care . . . بِتَوفِيرِ ضَرُورِيَّاتِ الحَيَاةِ الأَسَاسِيَّةِ لِشَعبِهَا وَمِن بَينِهَا الرِّعَايَةُ الطِّبِّيَّةُ



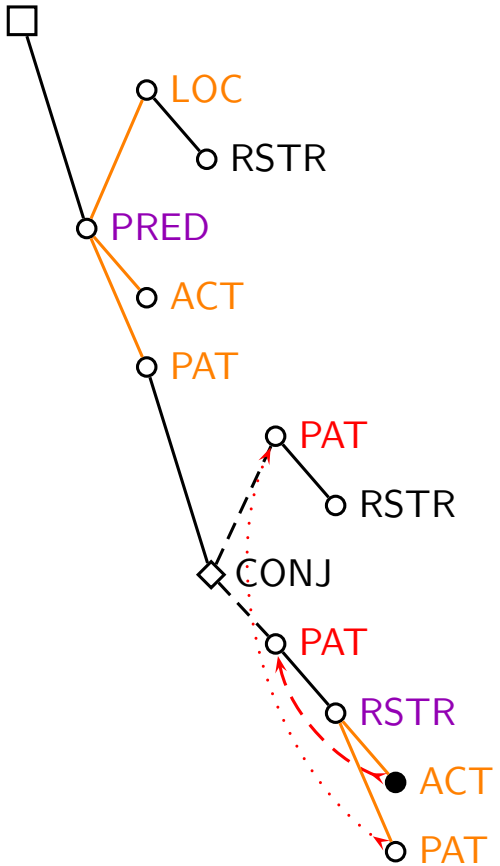
In the section on literature, the magazine presented the issue of the Arabic language and the dangers that threaten it. وَفِي مِلَفِّ الأَدَبِ طَرَحَتِ . . .



وَ *wa-*	and	C---------
فِي *fī*	in	P---------
مِلَفِّ *milaffi*	collection/file-of	N-------2R
اَلأَدَبِ *al-ʾadabi*	the-literature	N-------2D
طَرَحَت *ṭaraḥat*	it-presented	VP-A-3FS--
اَلمَجَلَّةُ *al-maǧallatu*	the-magazine	N-----FS1D
قَضِيَّةَ *qaḍīyata*	issue-of	N-----FS4R
اَللُّغَةِ *al-luġati*	the-language	N-----FS2D
اَلعَرَبِيَّةِ *al-ʿarabīyati*	the-Arabic	A-----FS2D
وَ *wa-*	and	C---------
اَلأَخطَارِ *al-ʾaḫṭāri*	the-dangers	N-------2D
اَلَّتِي *allatī*	that	SR----FS--
تُهَدِّدُ *tuhaddidu*	they-threaten	VIIA-3FS--
هَا *-hā*	it	S----3FS4-
. .	.	G---------

Tectogrammatics

Description of linguistic meaning in its semantic and pragmatic aspects.



B	ملف *milaff*	collection/file	Masc.Sing.Def
C	أدب *ʾadab*	literature	Masc.Sing.Def
B	طرح *ṭaraḥ*	to-present	Ind.Ant.Act
B	مجلة *mağallah*	magazine	Fem.Sing.Def
N	قضية *qaḍīyah*	issue	Fem.Sing.Def
N	لغة *luġah*	language	Fem.Sing.Def
N	عربي *ʿarabīy*	Arabic	Adjective
	و *wa*	and	Coordination
N	خطر *ḫaṭar*	danger	Masc.Plur.Def
N	هدد *haddad*	to-threaten	Ind.Sim.Act
B	هي *hiya*	it	PersPronoun
B	هي *hiya*	it	PersPronoun

Outline

Buckwalter Transliteration

يُولَدُ جَمِيعُ ٱلنَّاسِ أَحْرَارًا مُتَسَاوِينَ فِي ٱلْكَرَامَةِ وَٱلْحُقُوقِ. وَقَدْ وُهِبُوا عَقْلًا وَضَمِيرًا وَعَلَيْهِمْ أَنْ يُعَامِلَ بَعْضُهُمْ بَعْضًا بِرُوحِ ٱلْإِخَاءِ.

```
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 ArabTEX

يُولَدُ جَمِيعُ ٱلنَّاسِ أَحْرَارًا مُتَسَاوِينَ فِي ٱلْكَرَامَةِ وَٱلْحُقُوقِ. وَقَدْ وُهِبُوا عَقْلًا وَضَمِيرًا وَعَلَيْهِمْ أَنْ يُعَامِلَ بَعْضُهُمْ بَعْضًا بِرُوحِ ٱلْإِخَاءِ.

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

Yūladu ğamīʿu ’n-nāsi ʾaḥrāran mutasāwīna fī ’l-karāmati wa-’l-ḥuqūqi. Wa-qad wuhibū ʿaqlan wa-ḍamīran wa-ʿalayhim ʾan yuʿāmila baʿḍuhum baʿḍan bi-rūḥi ’l-ʾiẖāʾi.

```
\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 ← بِرُوحِ آلْإِخَاءِ ← 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

- 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/`