Yet Another Intro to Arabic NLP

Arabic Natural Language Processing

Otakar Smrž

Institute of Formal and Applied Linguistics
Charles University in Prague

Department of Middle Eastern Studies
University of West Bohemia in Pilsen

Autumn 2005

Yet Another Introduction to Arabic NLP

Arabic Natural Language Processing

Otakar Smrž

Institute of Formal and Applied Linguistics
Charles University in Prague

Department of Middle Eastern Studies
University of West Bohemia in Pilsen

Autumn 2005

This is the series of lecture notes to the course on Arabic Natural Language Processing taught in the Winter Term of 2005 at the Department of Middle Eastern Studies, Faculty of Philosophy, University of West Bohemia in Pilsen.

Lecturer Otakar Smrž <otakar.smrz@mff.cuni.cz>
Website http://ufal.mff.cuni.cz/~smrz/ANLP/

Interest of the Course

Natural Language Processing application of engineering to the problems of human languages

Computational Linguistics study of linguistic problems using computational methods :)

We *will not* do much of the crucial theories behind all that — theory of computation, logic, theoretical linguistics. We *will* explore how dealing with the natural language, esp. Arabic, is implemented today and what other solutions we can expect and contribute to.

Lecture Topics

1. Encodings, character sets, fonts, transliterations
   - Unicode, UTF-8, CP-1256, ...
   - Buckwalter transliteration
   - Meta-encodings of ArabTEX
   - Encode::Arabic
2. Data formats, markup, text processing and rendering
   - Plain text versus binary data
   - XML, HTML and LATEX documents
   - Writing in ArabTEX and MS Word
   - Data re-use, transcription, advanced sorting

Lecture Topics

1. Encodings, character sets, fonts, transliterations
   - Unicode, UTF-8, CP-1256, ...
   - Buckwalter transliteration
   - Meta-encodings of ArabTEX
   - Encode::Arabic
2. Data formats, markup, text processing and rendering
   - Plain text versus binary data
   - XML, HTML and LATEX documents
   - Writing in ArabTEX and MS Word
   - Data re-use, transcription, advanced sorting

Lecture Topics

1. Encodings, character sets, fonts, transliterations
   - Unicode, UTF-8, CP-1256, ...
   - Buckwalter transliteration
   - Meta-encodings of ArabTEX
   - Encode::Arabic
2. Data formats, markup, text processing and rendering
   - Plain text versus binary data
   - XML, HTML and LATEX documents
   - Writing in ArabTEX and MS Word
   - Data re-use, transcription, advanced sorting

Lecture Topics

1. Encodings, character sets, fonts, transliterations
   - Unicode, UTF-8, CP-1256, ...
   - Buckwalter transliteration
   - Meta-encodings of ArabTEX
   - Encode::Arabic
2. Data formats, markup, text processing and rendering
   - Plain text versus binary data
   - XML, HTML and LATEX documents
   - Writing in ArabTEX and MS Word
   - Data re-use, transcription, advanced sorting

Lecture Topics

1. Encodings, character sets, fonts, transliterations
   - Unicode, UTF-8, CP-1256, ...
   - Buckwalter transliteration
   - Meta-encodings of ArabTEX
   - Encode::Arabic
2. Data formats, markup, text processing and rendering
   - Plain text versus binary data
   - XML, HTML and LATEX documents
   - Writing in ArabTEX and MS Word
   - Data re-use, transcription, advanced sorting

Lecture Topics

1. Encodings, character sets, fonts, transliterations
   - Unicode, UTF-8, CP-1256, ...
   - Buckwalter transliteration
   - Meta-encodings of ArabTEX
   - Encode::Arabic
2. Data formats, markup, text processing and rendering
   - Plain text versus binary data
   - XML, HTML and LATEX documents
   - Writing in ArabTEX and MS Word
   - Data re-use, transcription, advanced sorting

Lecture Topics

1. Encodings, character sets, fonts, transliterations
   - Unicode, UTF-8, CP-1256, ...
   - Buckwalter transliteration
   - Meta-encodings of ArabTEX
   - Encode::Arabic
2. Data formats, markup, text processing and rendering
   - Plain text versus binary data
   - XML, HTML and LATEX documents
   - Writing in ArabTEX and MS Word
   - Data re-use, transcription, advanced sorting

Lecture Topics

1. Encodings, character sets, fonts, transliterations
   - Unicode, UTF-8, CP-1256, ...
   - Buckwalter transliteration
   - Meta-encodings of ArabTEX
   - Encode::Arabic
2. Data formats, markup, text processing and rendering
   - Plain text versus binary data
   - XML, HTML and LATEX documents
   - Writing in ArabTEX and MS Word
   - Data re-use, transcription, advanced sorting

Lecture Topics

1. Encodings, character sets, fonts, transliterations
   - Unicode, UTF-8, CP-1256, ...
   - Buckwalter transliteration
   - Meta-encodings of ArabTEX
   - Encode::Arabic
2. Data formats, markup, text processing and rendering
   - Plain text versus binary data
   - XML, HTML and LATEX documents
   - Writing in ArabTEX and MS Word
   - Data re-use, transcription, advanced sorting

Lecture Topics

1. Encodings, character sets, fonts, transliterations
   - Unicode, UTF-8, CP-1256, ...
   - Buckwalter transliteration
   - Meta-encodings of ArabTEX
   - Encode::Arabic
2. Data formats, markup, text processing and rendering
   - Plain text versus binary data
   - XML, HTML and LATEX documents
   - Writing in ArabTEX and MS Word
   - Data re-use, transcription, advanced sorting

Lecture Topics

3. Modeling of Arabic morphology, its various applications
   - Buckwalter morphological analyzer
   - Xerox finite-state morphology
   - Functional morphology
4. Lexicography and the other uses of linguistic corpora
   - Review of recent development and resources
5. Morphological and syntactic annotation projects
   - Prague Arabic Dependency Treebank
   - Penn Arabic Treebank

Lecture Topics

3. Modeling of Arabic morphology, its various applications
   - Buckwalter morphological analyzer
   - Xerox finite-state morphology
   - Functional morphology
4. Lexicography and the other uses of linguistic corpora
   - Review of recent development and resources
5. Morphological and syntactic annotation projects
   - Prague Arabic Dependency Treebank
   - Penn Arabic Treebank

Lecture Topics

3. Modeling of Arabic morphology, its various applications
   - Buckwalter morphological analyzer
   - Xerox finite-state morphology
   - Functional morphology
4. Lexicography and the other uses of linguistic corpora
   - Review of recent development and resources
5. Morphological and syntactic annotation projects
   - Prague Arabic Dependency Treebank
   - Penn Arabic Treebank

Lecture Topics

3. Modeling of Arabic morphology, its various applications
   - Buckwalter morphological analyzer
   - Xerox finite-state morphology
   - Functional morphology
4. Lexicography and the other uses of linguistic corpora
   - Review of recent development and resources
5. Morphological and syntactic annotation projects
   - Prague Arabic Dependency Treebank
   - Penn Arabic Treebank

Lecture Topics

3. Modeling of Arabic morphology, its various applications
   - Buckwalter morphological analyzer
   - Xerox finite-state morphology
   - Functional morphology
4. Lexicography and the other uses of linguistic corpora
   - Review of recent development and resources
5. Morphological and syntactic annotation projects
   - Prague Arabic Dependency Treebank
   - Penn Arabic Treebank

Lecture Topics

3. Modeling of Arabic morphology, its various applications
   - Buckwalter morphological analyzer
   - Xerox finite-state morphology
   - Functional morphology
4. Lexicography and the other uses of linguistic corpora
   - Review of recent development and resources
5. Morphological and syntactic annotation projects
   - Prague Arabic Dependency Treebank
   - Penn Arabic Treebank

Lecture Topics

3. Modeling of Arabic morphology, its various applications
   - Buckwalter morphological analyzer
   - Xerox finite-state morphology
   - Functional morphology
4. Lexicography and the other uses of linguistic corpora
   - Review of recent development and resources
5. Morphological and syntactic annotation projects
   - Prague Arabic Dependency Treebank
   - Penn Arabic Treebank

Lecture Topics

3. Modeling of Arabic morphology, its various applications
   - Buckwalter morphological analyzer
   - Xerox finite-state morphology
   - Functional morphology
4. Lexicography and the other uses of linguistic corpora
   - Review of recent development and resources
5. Morphological and syntactic annotation projects
   - Prague Arabic Dependency Treebank
   - Penn Arabic Treebank

Lecture Topics

3. Modeling of Arabic morphology, its various applications
   - Buckwalter morphological analyzer
   - Xerox finite-state morphology
   - Functional morphology
4. Lexicography and the other uses of linguistic corpora
   - Review of recent development and resources
5. Morphological and syntactic annotation projects
   - Prague Arabic Dependency Treebank
   - Penn Arabic Treebank

Other Information

ANLP Homepage > Software Installation Guide

`http://ufal.mff.cuni.cz/~smrz/ANLP/.php?html=tools`

ANLP Homepage > Bibliography and References

`http://ufal.mff.cuni.cz/~smrz/ANLP/.php?html=links`

Encodings and Transliterations

Arabic Natural Language Processing

Otakar Smrž

Institute of Formal and Applied Linguistics
Charles University in Prague

Department of Middle Eastern Studies
University of West Bohemia in Pilsen

October 6, 2005

Where to Begin . . .

Romanization, Transcription and Transliteration *by Ken Beesley*

`http://www.xrce.xerox.com/competencies/content-analysis/arabic/info/romanization.html`

orthography conventions for representing a language with some set of symbols or character set

transcription alternative (phonetic, morphological) representation of the language, possibly romanization

transliteration orthography with carefully substituted symbols, yet preserving the original conventions

encoding transliteration mapping the orthographical symbols into numbers interpreted as characters or bytes

Where to Begin . . .

Romanization, Transcription and Transliteration *by Ken Beesley*

`http://www.xrce.xerox.com/competencies/content-analysis/arabic/info/romanization.html`

orthography conventions for representing a language with some set of symbols or character set

transcription alternative (phonetic, morphological) representation of the language, possibly romanization

transliteration orthography with carefully substituted symbols, yet preserving the original conventions

encoding transliteration mapping the orthographical symbols into numbers interpreted as characters or bytes

Where to Begin . . .

Romanization, Transcription and Transliteration *by Ken Beesley*

`http://www.xrce.xerox.com/competencies/content-analysis/arabic/info/romanization.html`

orthography conventions for representing a language with some set of symbols or character set

transcription alternative (phonetic, morphological) representation of the language, possibly romanization

transliteration orthography with carefully substituted symbols, yet preserving the original conventions

encoding transliteration mapping the orthographical symbols into numbers interpreted as characters or bytes

Where to Begin . . .

Romanization, Transcription and Transliteration *by Ken Beesley*

`http://www.xrce.xerox.com/competencies/content-analysis/arabic/info/romanization.html`

orthography conventions for representing a language with some set of symbols or character set

transcription alternative (phonetic, morphological) representation of the language, possibly romanization

transliteration orthography with carefully substituted symbols, yet preserving the original conventions

encoding transliteration mapping the orthographical symbols into numbers interpreted as characters or bytes

Bits of Information

Computers represent any information only through sequences of bits. Eight-tuples of bits form bytes. Every bit can only distinguish two values, symbols of a binary numeral system. Their series are identified as numbers, which may also encode, or be viewed as, characters.

Wikipedia > Binary numeral system

`http://en.wikipedia.org/wiki/Binary_numeral_system`

Wikipedia > ASCII

`http://en.wikipedia.org/wiki/ASCII`

Unicode in Its Words

Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language.

Před vznikem Unicode existovaly stovky rozdílných kódovacích systémů pro přiřazování těchto čísel.

وَلَم يُوجَد نِظَامُ تَشفِيرٍ وَاحِدٌ يَحتَوِي عَلَى جَمِيعِ المَحَارِفِ الضَّرُورِيَّةِ.

Wa-lam yūğad niẓāmu tašfīrin wāḥidun yaḥtawī ʿalā ğamīʿi ’l-maḥārifi ’ḍ-ḍarūrīyati.

Unicode

`http://www.unicode.org/`

Unicode Arabic

The identifiers that are assigned to characters are called code points.

- Arabic U+0600–U+06FF
- Arabic Supplement U+0750–U+077F
- Arabic Presentation Forms-A U+FB50–U+FDFF
- Arabic Presentation Forms-B U+FE70–U+FEFF

Unicode also specifies algorithms for contextual and bidirectional rendering of the characters, while Universal Character Set does not. (more on that later)

Wikipedia > Universal Character Set

`http://en.wikipedia.org/wiki/Universal_Character_Set`

Encodings of Unicode

Furthermore, the numbers of code points need to be mapped into some well-defined sequences of bytes. There are various ways to do that, and the mappings are called Unicode Transformation Formats.

Unicode Code Point	UTF-8 Byte Sequence		Character
U+0639 0000 0110 0011 1001	0xD8 1101 1000	0xB9 1011 1001	Arabic ع
U+004C 0000 0000 0100 1100	0x4C 0100 1100		Latin L

Wikipedia > UTF-8

http://en.wikipedia.org/wiki/UTF-8

Encodings of Unicode

Furthermore, the numbers of code points need to be mapped into some well-defined sequences of bytes. There are various ways to do that, and the mappings are called Unicode Transformation Formats.

Unicode Code Point	UTF-8 Byte Sequence		Character
U+0639 0000 0110 0011 1001	0xD8 1101 1000	0xB9 1011 1001	Arabic ع
U+004C 0000 0000 0100 1100	0x4C 0100 1100		Latin L

Wikipedia > UTF-8

http://en.wikipedia.org/wiki/UTF-8

Buckwalter Transliteration

ء	’	ء	*d*	d	د	*ḍ*	D	ض	*k*	k	ك
b	b	ب	*ḏ*	*	ذ	*ṭ*	T	ط	*l*	l	ل
t	t	ت	*r*	r	ر	*ẓ*	Z	ظ	*m*	m	م
ṯ	v	ث	*z*	z	ز	‘	E	ع	*n*	n	ن
ǧ	j	ج	*s*	s	س	*ġ*	g	غ	*h*	h	ه
ḥ	H	ح	*š*	$	ش	*f*	f	ف	*w*	w	و
ḫ	x	خ	*ṣ*	S	ص	*q*	q	ق	*y*	y	ي
ā	A	ا	*p*	P	پ	*v*	V	ڤ	ء	W	ؤ
ء	O	أ	*č*	J	چ	*g*	G	گ	ء	}	ئ
ء	I	إ	*ž*	R	ژ	*h*	p	ة	*ā*	Y	ى

Buckwalter Transliteration

'	'	ء	*d*	d	د	*ḍ*	D	ض	*k*	k	ك
b	b	ب	*ḏ*	*	ذ	*ṭ*	T	ط	*l*	l	ل
t	t	ت	*r*	r	ر	*ẓ*	Z	ظ	*m*	m	م
ṯ	v	ث	*z*	z	ز	‘	E	ع	*n*	n	ن
ǧ	j	ج	*s*	s	س	*ġ*	g	غ	*h*	h	ه
ḥ	H	ح	*š*	$	ش	*f*	f	ف	*w*	w	و
ḫ	x	خ	*ṣ*	S	ص	*q*	q	ق	*y*	y	ي
ā	A	ا	*p*	P	پ	*v*	V	ڤ	'	W	ؤ
'	O	أ	*č*	J	چ	*g*	G	گ	'	}	ئ
'	I	إ	*ž*	R	ژ	*h*	p	ة	*ā*	Y	ى

Buckwalter Transliteration

‘	’	ء	*d*	d	د	*ḍ*	D	ض	*k*	k	ك
b	b	ب	*ḏ*	*	ذ	*ṭ*	T	ط	*l*	l	ل
t	t	ت	*r*	r	ر	*ẓ*	Z	ظ	*m*	m	م
ṯ	v	ث	*z*	z	ز	‘	E	ع	*n*	n	ن
ǧ	j	ج	*s*	s	س	*ġ*	g	غ	*h*	h	ه
ḥ	H	ح	*š*	$	ش	*f*	f	ف	*w*	w	و
ḫ	x	خ	*ṣ*	S	ص	*q*	q	ق	*y*	y	ي
ā	A	ا	*p*	P	پ	*v*	V	ڤ	‘	W	ؤ
‘	O	أ	*č*	J	چ	*g*	G	گ	‘	}	ئ
‘	I	إ	*ž*	R	ژ	*h*	p	ة	*ā*	Y	ى

Buckwalter Transliteration

'	'	ء	d	d	د	ḍ	D	ض	k	k	ك
b	b	ب	ḏ	*	ذ	ṭ	T	ط	l	l	ل
t	t	ت	r	r	ر	ẓ	Z	ظ	m	m	م
ṯ	v	ث	z	z	ز	‘	E	ع	n	n	ن
ǧ	j	ج	s	s	س	ġ	g	غ	h	h	ه
ḥ	H	ح	š	$	ش	f	f	ف	w	w	و
ḫ	x	خ	ṣ	S	ص	q	q	ق	y	y	ي
ā	A	ا	p	P	پ	v	V	ڤ	'	W	ؤ
'	O	أ	č	J	چ	g	G	گ	'	}	ئ
'	I	إ	ž	R	ژ	h	p	ة	ā	Y	ى

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.
```

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 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

ء	’	ء	*d*	d	د	*ḍ*	.d	ض	*k*	k	ك
b	b	ب	*ḏ*	_d	ذ	*ṭ*	.t	ط	*l*	l	ل
t	t	ت	*r*	r	ر	*ẓ*	.z	ظ	*m*	m	م
ṯ	_t	ث	*z*	z	ز	‘	‘	ع	*n*	n	ن
ǧ	^g	ج	*s*	s	س	*ġ*	.g	غ	*h*	h	ه
ḥ	.h	ح	*š*	^s	ش	*f*	f	ف	*w*	w	و
ḫ	_h	خ	*ṣ*	.s	ص	*q*	q	ق	*y*	y	ي
ā	A	ا	*p*	p	پ	*v*	v	ڤ	ء	’	ؤ
ء	’	أ	*č*	^c	چ	*g*	g	گ	ء	’	ئ
ء	’	إ	*ž*	^z	ژ	*h*	T	ة	*ā*	Y	ى

Notation of ArabTEX

ɔ	’	ء	*d*	d	د	*ḍ*	.d	ض	*k*	k	ك
b	b	ب	*ḏ*	_d	ذ	*ṭ*	.t	ط	*l*	l	ل
t	t	ت	*r*	r	ر	*ẓ*	.z	ظ	*m*	m	م
ṯ	_t	ث	*z*	z	ز	ʿ	‘	ع	*n*	n	ن
ǧ	^g	ج	*s*	s	س	*ġ*	.g	غ	*h*	h	ه
ḥ	.h	ح	*š*	^s	ش	*f*	f	ف	*w*	w	و
ḫ	_h	خ	*ṣ*	.s	ص	*q*	q	ق	*y*	y	ي
ā	A	ا	*p*	p	پ	*v*	v	ڤ	ɔ	’	ؤ
ɔ	’	أ	*č*	^c	چ	*g*	g	گ	ɔ	’	ئ
ɔ	’	إ	*ž*	^z	ژ	*h*	T	ة	*ā*	Y	ى

Notation of ArabTEX

’	’|	ء	*d*	d	د	*ḍ*	.d	ض	*k*	k	ك
b	b	ب	*ḏ*	_d	ذ	*ṭ*	.t	ط	*l*	l	ل
t	t	ت	*r*	r	ر	*ẓ*	.z	ظ	*m*	m	م
ṯ	_t	ث	*z*	z	ز	‘	‘	ع	*n*	n	ن
ǧ	^g	ج	*s*	s	س	*ġ*	.g	غ	*h*	h	ه
ḥ	.h	ح	*š*	^s	ش	*f*	f	ف	*w*	w	و
ḫ	_h	خ	*ṣ*	.s	ص	*q*	q	ق	*y*	y	ي
ā	A	ا	*p*	p	پ	*v*	v	ڤ	’	’w	ؤ
’	’a	أ	*č*	^c	چ	*g*	g	گ	’	’y	ئ
’	’i	إ	*ž*	^z	ژ	*h*	T	ة	*ā*	Y	ى

Notation of ArabTEX

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

```
\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 ArabTEX

يولد جميع النّاس 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 ArabTEX

Yūladu ǧamīʿu ’n-nāsi ʾaḥrāran mutasāwīna fī ’l-karāmati wa-’l-ḥuqūqi. Waqad wuhibū ʿaqlan wa-ḍamīran wa-ʿalayhim ʾan yuʿāmila baʿḍuhum baʿḍan bi-rūḥi ’l-ʾiǧāʾi.

```
\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 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 ǧamīʿu ’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.
```

Other Resources

Unicode > Code Charts

http://www.unicode.org/charts/

Buckwalter Transliteration

http://www.qamus.org/transliteration.htm

ArabTEX User Manual

http://129.69.218.213/arabtex/html/arabdoc.pdf

Encode::Arabic Online Interface

http://ufal.mff.cuni.cz/~smrz/Encode/Arabic/

Omniglot > Arabic Script

http://www.omniglot.com/writing/arabic.htm

Documents and Data — Using LaTeX

Arabic Natural Language Processing

Otakar Smrž

Institute of Formal and Applied Linguistics
Charles University in Prague

Department of Middle Eastern Studies
University of West Bohemia in Pilsen

October 13 & 20, 2005

Preliminaries

document collection of data, logical rather than physical (cf. file)

data information in the form suited for computer processing

We will explore how re-usable documents can be defined with the help of markup and a little bit of programming. We will separate the information content of our documents from their visual form. We will need LaTeX and ArabTeX to create one from the other.

ArabTeX User Manual Version 4.00 *by Klaus Lagally*

`http://129.69.218.213/arabtex/doc/arabdoc.pdf`

Expressing Meaning

Information requires structure. Explicit annotation of structure to a given expression in text or data is done through abstract markup. The interpretation of the markup in turn determines the meaning of the expression.

Example: structuring this very slide using the notation of LaTeX

```
\begin{slide}
    \title{Expressing Meaning}

    Information requires \alert{structure}.
    Explicit \alert{annotation} of structure ...
\end{slide}
```

Writing in LaTeX

Note the notation for the markup, esp. \ { }, and the pairing and possible nesting of \begin{...} and \end{...}.

Example: elementary document in LaTeX

```
\documentclass{article}

\begin{document}

    This is the first paragraph of my document.

    This is the second paragraph, completing the
    document for now.  There is not much to it.

\end{document}
```

Including ArabTEX

We can use ready-made packages to extend or change the interpretation of our document. We can include الخَطّ العَرَبِيّ, of course.

Example: insertions of the right-to-left script

```
\documentclass{article}

\usepackage{arabtex}

\begin{document}

    This is the first and only paragraph of my document
    \<wa-fIhA _ha.t.tuN ‘arabIyuN> \RL{’aw h_aka_dA}.

\end{document}
```

Flashcards!

`\documentclass{flashcards}` provides us with another functionality — it will typeset flashcards with vocabulary for us, with some predefined style, which we now modify explicitly to get the Arabic side both in the orthography and in the transcription.

Example: this is the document part of our document

```
\begin{flashcard}{sun beams}
    \RL{’a^si‘‘aTu a^s-^samsi}  % this text is a comment
    \\ \arabfalse\transtrue
    \RL{’a^si‘‘aTu a^s-^samsi}
\end{flashcard}

\begin{flashcard}{official measures}
    \RL{’i^grA’AtuN rasmIyaTuN}  \\ \arabfalse\transtrue
    \RL{’i^grA’AtuN rasmIyaTuN}
\end{flashcard}
```

Funny Morphology and MorphoTrees

Arabic Natural Language Processing

Otakar Smrž

Institute of Formal and Applied Linguistics
Charles University in Prague

Department of Middle Eastern Studies
University of West Bohemia in Pilsen

November 24, 2005

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 **strings** consist of up to four syntactic tokens

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 **strings** consist of up to four syntactic tokens

- 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

Functional Approximation

The underlying morphological engine for both the Penn Arabic Treebank and the Prague Arabic Dependency Treebank is the Buckwalter Arabic Morphological Analyzer. While PATB adopts the analyses in their original format, the PADT annotations take place on approximations of Functional Arabic Morphology, a novel theoretical and computational model, organized into MorphoTrees.

Functional Approximation

The underlying morphological engine for both the Penn Arabic Treebank and the Prague Arabic Dependency Treebank is the Buckwalter Arabic Morphological Analyzer. While PATB adopts the analyses in their original format, the PADT annotations take place on approximations of Functional Arabic Morphology, a novel theoretical and computational model, organized into MorphoTrees.

With respect to the linguistic view and the architecture of the software that we develop, we unify the format of the morphological data by converting all the Parts of PATB into the approximation, which is done in two steps: (a) the morphs of the original input strings are re-grouped to form tokens (b) the corresponding sequences of tags are mapped into the fixed-width positional notation of PADT.

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

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
سيخبرهم		F---------	FUT	*sa-*	will
		VIIA-3MS--	IV3MS+IV+IVSUFF_MOOD:I	*yu-ḫbir-u*	he-notify
		S----3MP4-	IVSUFF_DO:3MS	*-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

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
سيخبرهم		F---------	FUT	*sa-*	will
		VIIA-3MS--	IV3MS+IV+IVSUFF_MOOD:I	*yu-ḫbir-u*	he-notify
		S----3MP4-	IVSUFF_DO:3MS	*-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

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

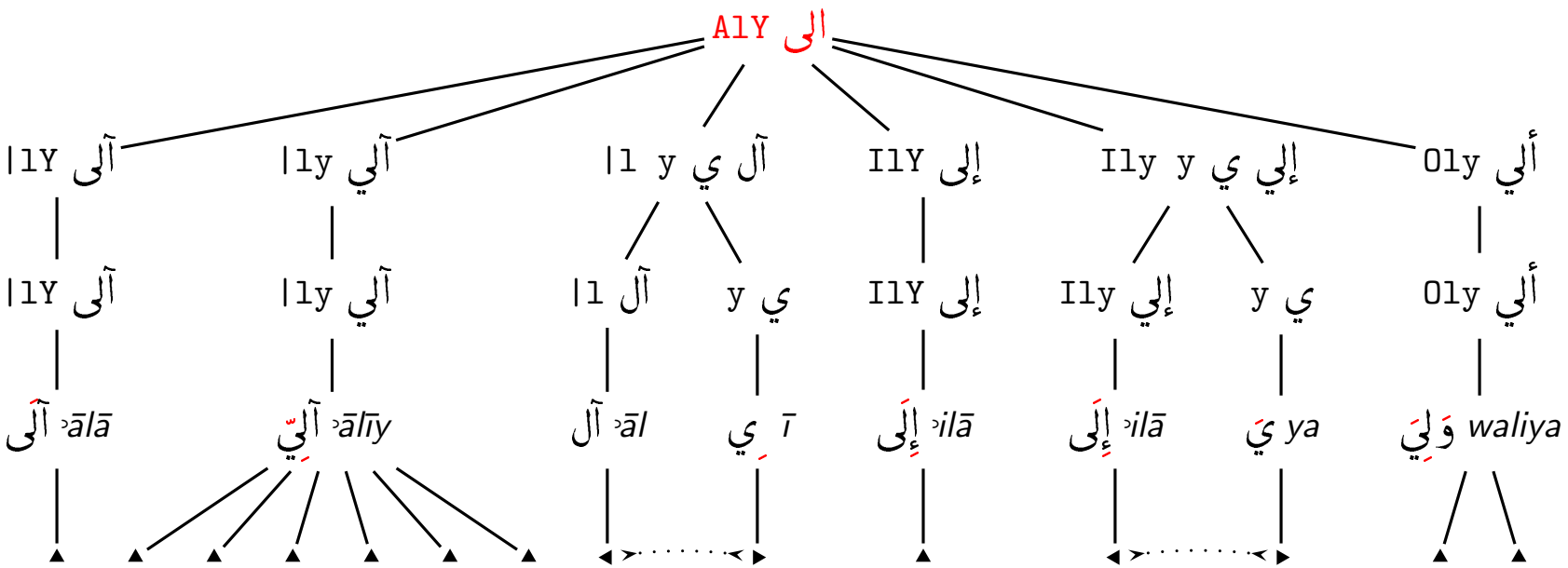
String	Token	Token Tag	Buckwalter's M-Tags	Token Form	Token Gloss
سيخبرهم		F---------	FUT	*sa-*	will
		VIIA-3MS--	IV3MS+IV+IVSUFF_MOOD:I	*yu-ḫbir-u*	he-notify
		S----3MP4-	IVSUFF_DO:3MS	*-hum*	them
بذلك		P---------	PREP	*bi-*	about/by
		SD----MS--	DEM_PRON_MS	*ḏālika*	that
عن		P---------	PREP	*ʿan*	by/about
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 vs. Lists

Suppose you have a list of morphological analyses [the next slide] for a given input string . . .

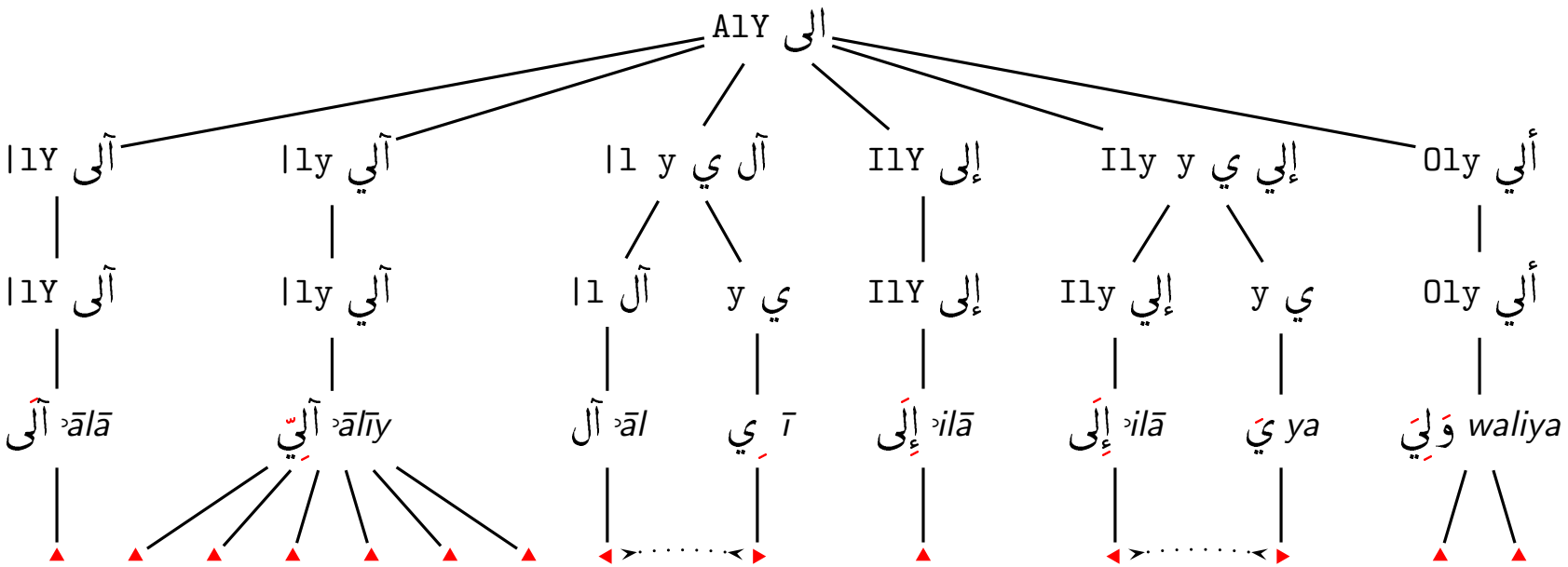
MorphoTrees vs. Lists

. . . organize them into a hierarchy with the string as its root



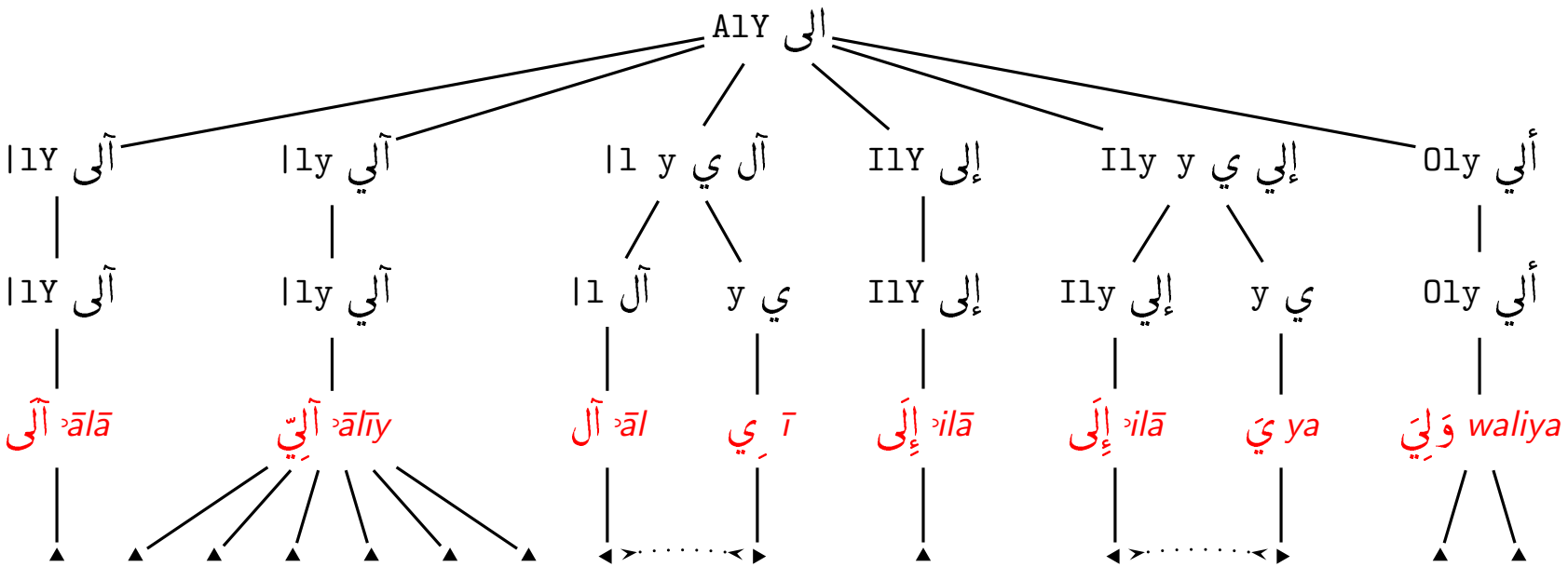
MorphoTrees vs. Lists

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



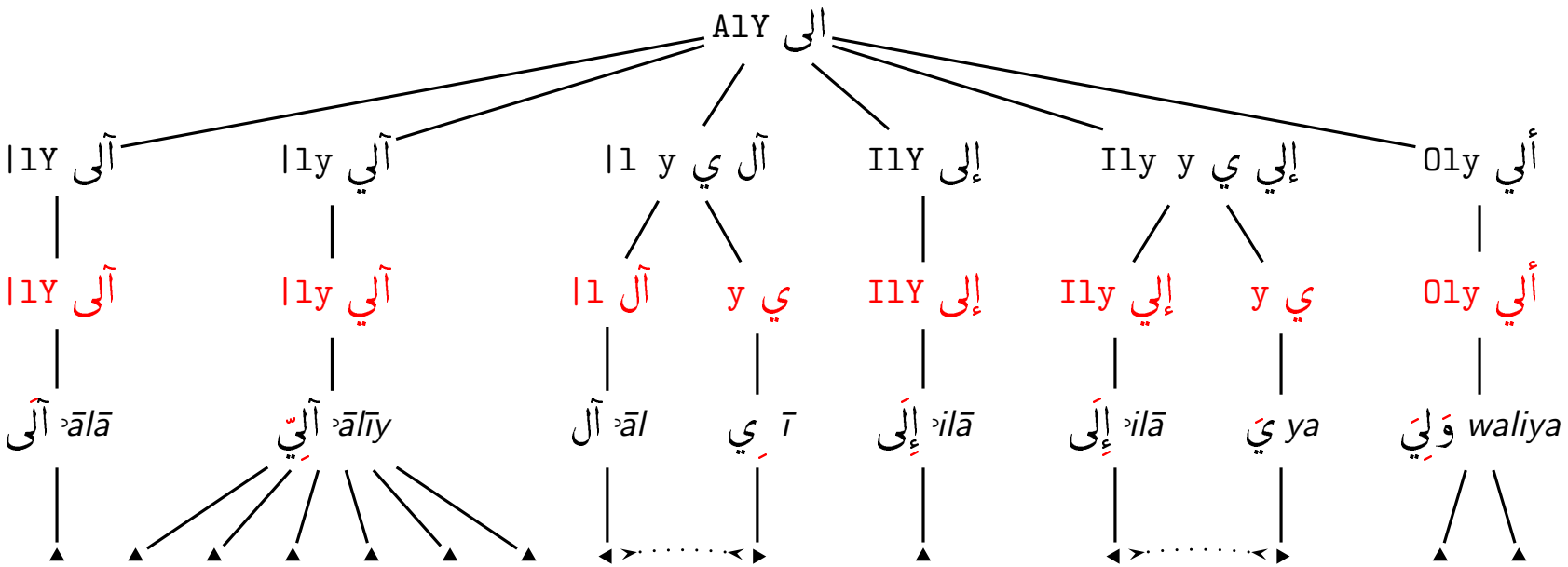
MorphoTrees vs. Lists

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



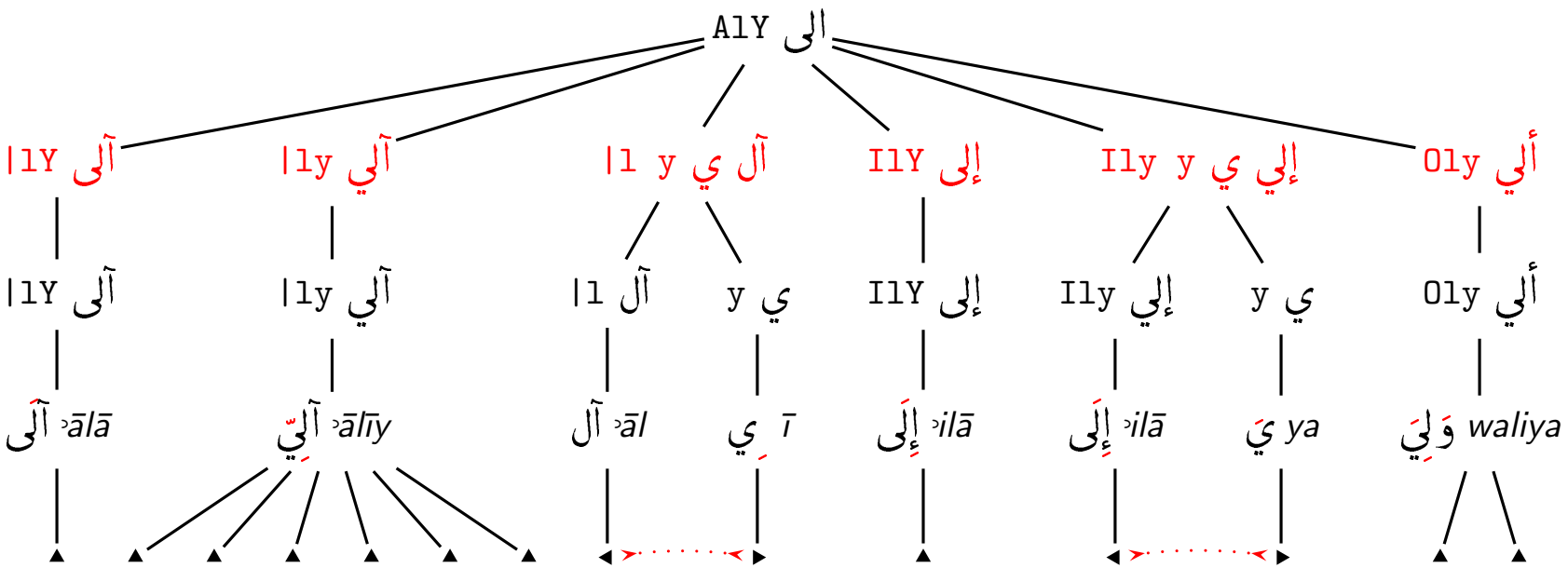
MorphoTrees vs. Lists

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



MorphoTrees vs. Lists

. . . organize them 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:



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.]

List of the morphological analyses of the input string AlY الى with the quasi-functional token tags derived from the original Buckwalter's tags.

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.]

String Tag VISA-1-S--

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.]

String Tag N--------RS----1-S2-

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.]

Ambiguity Classes

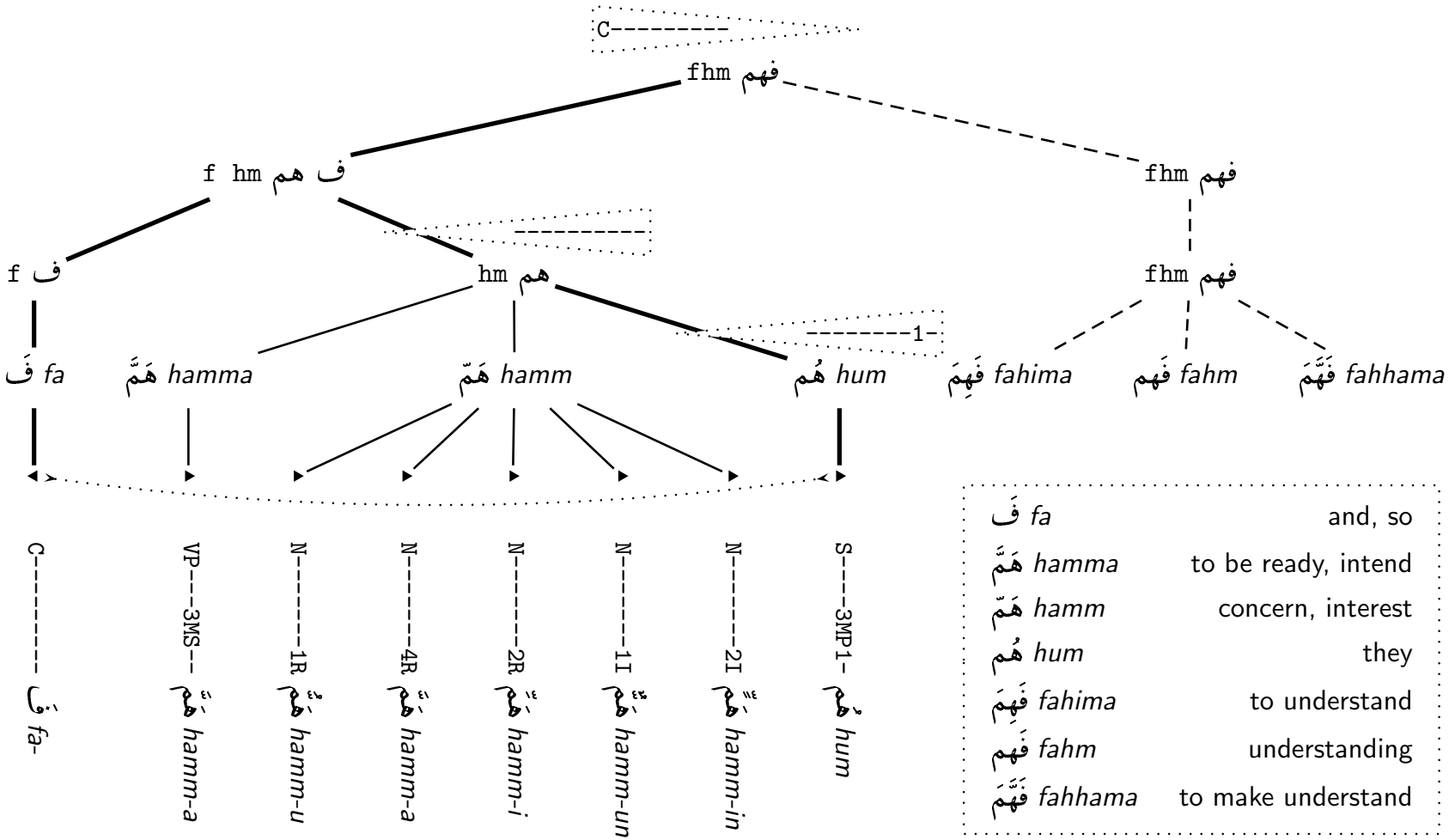
```
VP-A-3MS-----------------------------
A-I- ---1RS    1 S2
NIS  1  2I
P       4
```

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.]

Ambiguity Classes

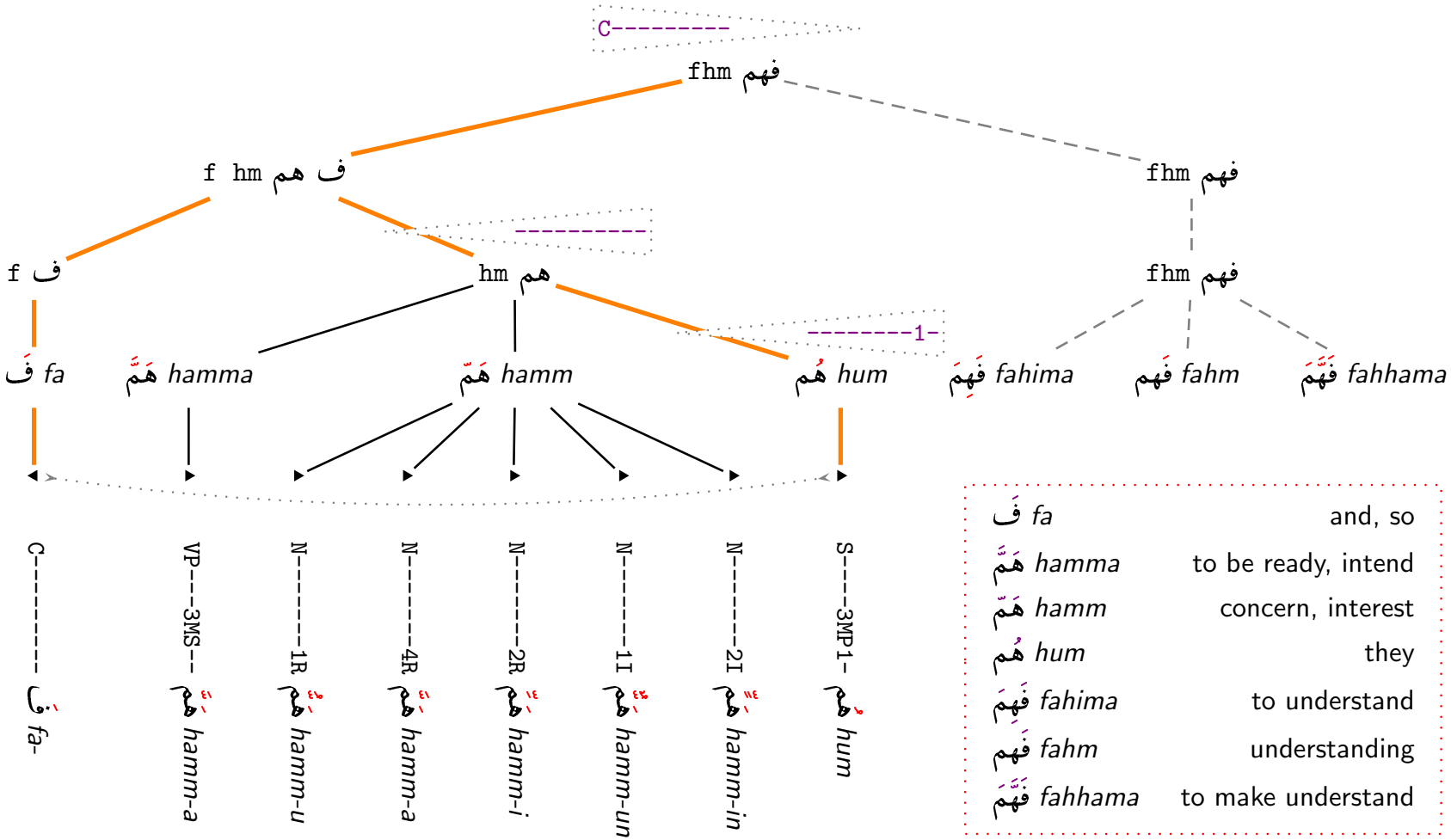
```
V         -
AP-   3   1-
N-IA -MS2R-     - --
PIS--1--4IS----1-S2--------------------
```

MorphoTrees with Restrictions



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

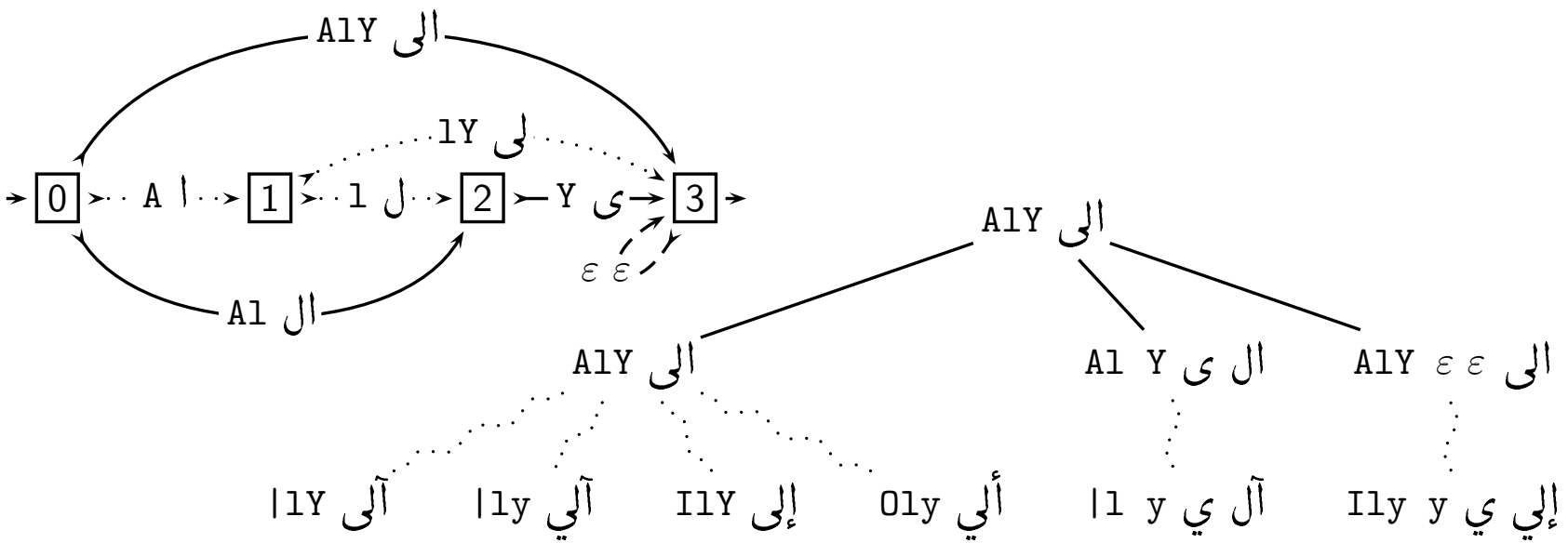
MorphoTrees with Restrictions



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

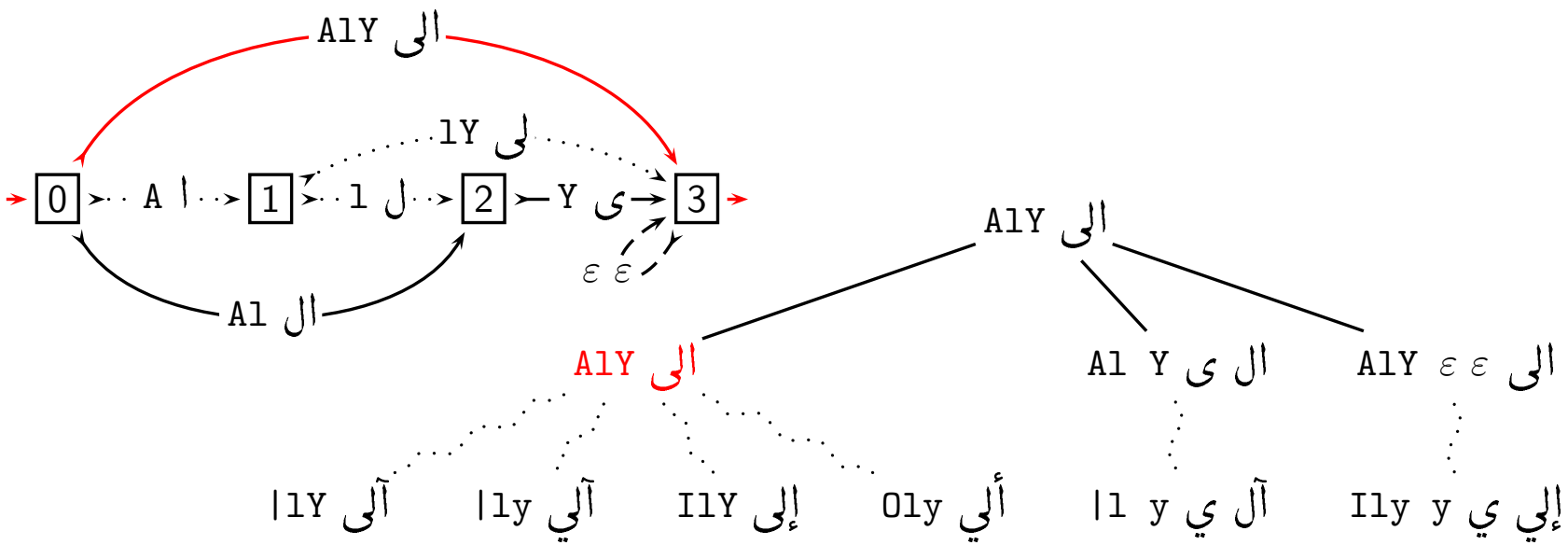
Tknz versus Tknz++

We introduce two measures for tokenization. Tknz is close to the evaluations that only check the partitioning determined by finding token boundaries between the characters of the original string, and do not, unlike Tknz++, require the tokenization to faithfully reconstruct the canonical non-vocalized forms of tokens, as is the case in MorphoTrees.



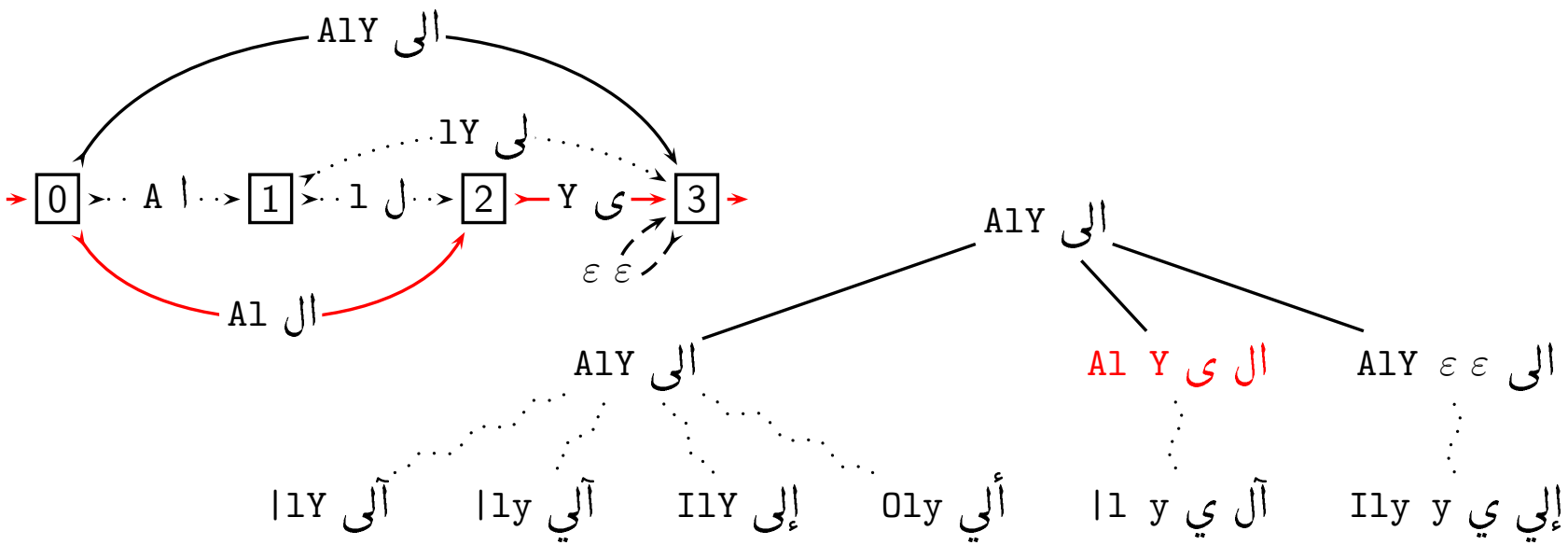
Tknz versus Tknz++

We introduce two measures for tokenization. Tknz is close to the evaluations that only check the partitioning determined by finding token boundaries between the characters of the original string, and do not, unlike Tknz++, require the tokenization to faithfully reconstruct the canonical non-vocalized forms of tokens, as is the case in MorphoTrees.



Tknz versus Tknz++

We introduce two measures for tokenization. Tknz is close to the evaluations that only check the partitioning determined by finding token boundaries between the characters of the original string, and do not, unlike Tknz++, require the tokenization to faithfully reconstruct the canonical non-vocalized forms of tokens, as is the case in MorphoTrees.



Tknz versus Tknz++

We introduce two measures for tokenization. Tknz is close to the evaluations that only check the partitioning determined by finding token boundaries between the characters of the original string, and do not, unlike Tknz++, require the tokenization to faithfully reconstruct the canonical non-vocalized forms of tokens, as is the case in MorphoTrees.