Transitions all the Way Down: From Characters to Full Document Annotation in One System

Yuval Pinter
Ben-Gurion University

Working Groups 1 & 3

Miryam de Lhoneux
KU Leuven

Motivation

Compositionality:

- characters form morphemes
- morphemes form words
- words form sentences

Not observable in text

#	form	lemma	pos	head	relation	morphosyntactic attributes
1-3	BBTYM					
1	B		PREP	3	case	
2	H		DT	3	det	
3	BTYM	BYT	NOUN	0	root	Number=Plural;Def=Definite
4-5	HGDWLYM					
4	H		DT	5	det	
5	GDWLYM	GDWL	ADJ	3	amod	Number=Plural;Def=Definite

Syntactic fusion

Table 1: Desired tagged output for the Hebrew fragment 'BBTYM HGDWLYM'.

Conventional plural markers

בבתיהגדולים
BBTYM HGDWLYM
ba-batim *ha-gdolim*
in-(the-)house.pl the-big.pl

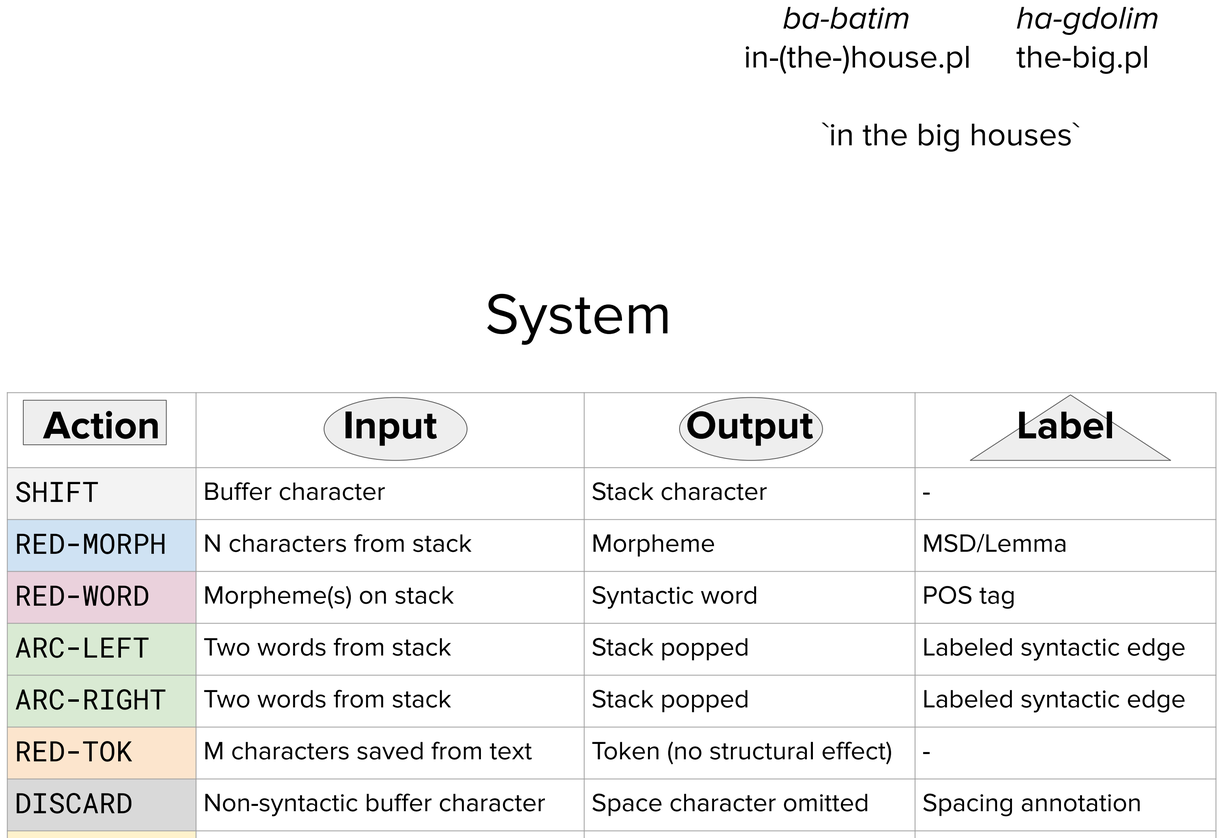
`in the big houses`

Walkthrough (1st token)

Action	Result	Notes
SHIFT	Character buffer → character node on stack	
RED-MORPH(null)	Morpheme node on stack	No MSD or lemma for this word
RED-WORD(PREP)	Word node on stack, POS tagged	
RED-MORPH(Def=Definite)	Morpheme node on stack, MSD stored	
RED-WORD(DT)	Word node on stack	Orthographically unrealized word
SHIFT	Character buffer → character node on stack	
SHIFT	Character buffer → character node on stack	
RED-MORPH(Lemma=`BYT`)	Morpheme node, lemma stored	Two character nodes combine
SHIFT	Character buffer → character node on stack	
SHIFT	Character buffer → character node on stack	
RED-MORPH(Number=Plur)	Morpheme node, MSD stored	Suffix carrying MSD
RED-WORD(NOUN)	Word node, POS tagged	Two morpheme nodes combine; stored lemma and MSDs tagged
ARC-LEFT(det)	Dep arc created, stack popped	
ARC-LEFT(case)	Dep arc created, stack popped	
RED-TOK	Token annotated	
DISCARD	Character buffer → trash	Space character omitted
[…]	[…]	[Next token(s) processed]
RED-SENT(null)	Sentence node on stack	Optional for discourse tagging

System

Action	Input	Output	Label
SHIFT	Buffer character	Stack character	-
RED-MORPH	N characters from stack	Morpheme	MSD/Lemma
RED-WORD	Morpheme(s) on stack	Syntactic word	POS tag
ARC-LEFT	Two words from stack	Stack popped	Labeled syntactic edge
ARC-RIGHT	Two words from stack	Stack popped	Labeled syntactic edge
RED-TOK	M characters saved from text	Token (no structural effect)	-
DISCARD	Non-syntactic buffer character	Space character omitted	Spacing annotation
RED-SENT	ROOT token from stack	Sentence token	Discourse annotation



Annotation

- Our main resource is **UD** (standard datasets & splits)
- Main oracle challenge - align morphemes to produce MSDs ← we wrote a script to extract morpheme segmentations before oracle construction, using **MorphyNet** annotations
  - Challenges that remain - infixes and other nonconcatenative morphological relations (possible remedy: methods from nonprojective parsing)
- Languages in our current development state: **English**, **Catalan**, **Swedish**, **Italian**, **Hungarian**, *[**Hebrew**, **Turkish** - w/o MorphyNet]*

Implementation