

NooJ'2008 Conference

Morphological grammars for standard Arabic tokenisation

Budapest, 8 – 10 June 2008

Slim MESFAR

LASELDI - Franche-Comté University – France

mesfarslim@yahoo.fr

Preliminary

Main characteristics of Arabic:

- Agglutinative word structure
 - The Arabic language is a strongly agglutinative language;
 - Most word forms in Arabic writings can correspond to a succession of :
 - one or more proclitics,
 - a radical,
 - an enclitic
- Non vocalization
- Lexical ambiguity
- Flexibility of word's order in sentences

Proclitics and Enclitics

10 proclitics + 13 enclitics

1 st position	2 nd position	3 rd position	4 th position
Question article “أ” (á - did/does)	Conjunctions : ف (fa) and و (wa)	- Prepositions : ب (bi – with), ل (li – for) and ك (ka – like) - Subjunctive or apocope particle : ل (li) - Future particle : س (sa) - Corroboration mark : لا (la)	The definite article “ال” (el – the)

} Possible proclitic positions

Their combination is governed by two types of constraints:

- Order relationship
- Compatibility rules

Possible concatenations in nouns

Proclitics				Radical	Enclitics
Question article	Conjunctions	Prepositions	Definite article	Noun	Personal pronoun
Question mark ا “a” (does or did)	Conjunctions و “wa” (and) or ف “fa” (then)	Prepositions : ب (bi – with), ل (li – for) et ك (ka – like)	The definite article “ال” (el – the)	Inflected form	First person (2)
					Second person (5)
					Third person (6)

→ Among 448 statistically possible concatenations, only 52 give valid forms

→ need to apply some validity constraints

Lexical analysis

For the word form « أوحل », we can have 3 potential analyses :

✓ 1st analysis :

1 st proclitic	2 nd proclitic	Inflected form
أ (á) – question mark	و (wa) – coordination conjunction	حَلَّ : preterit verb (<i>ħall</i> – resolve) or a noun (<i>ħall</i> – solution)

✓ 2nd analysis :

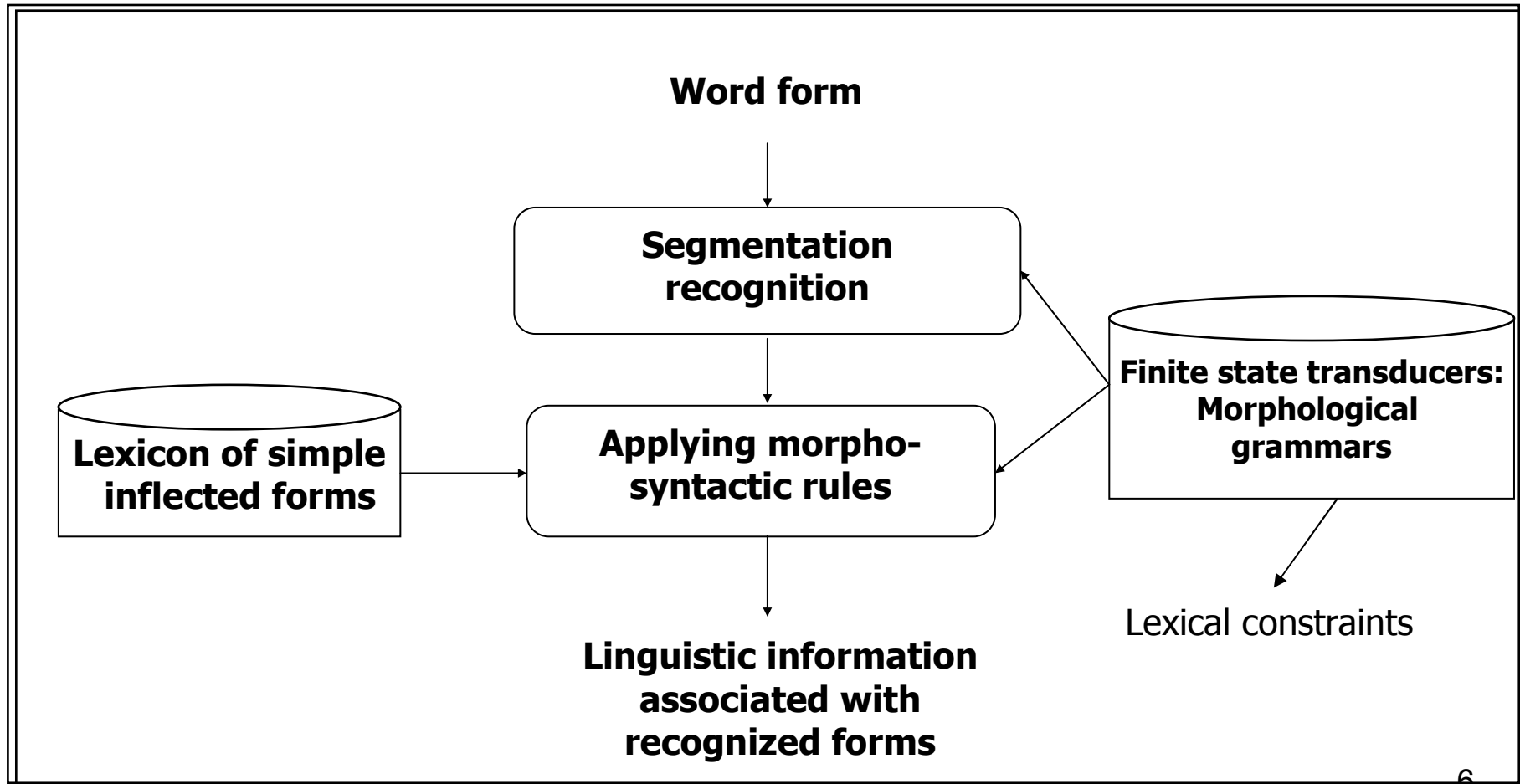
Proclitic	Inflected form
أ (á) – question mark	وَحَلَ : preterit verb (<i>waħal</i> – jam) or a noun (<i>waħl</i> – mud)

✓ 3rd analysis :

Inflected form
أوَحَلَ – preterit verb (<i>áawħala</i> – throw in the mud)

→ Need of a morphological analyser : separates and identifies the input word morphemes + labels them with sufficient information.

Morphological analysis



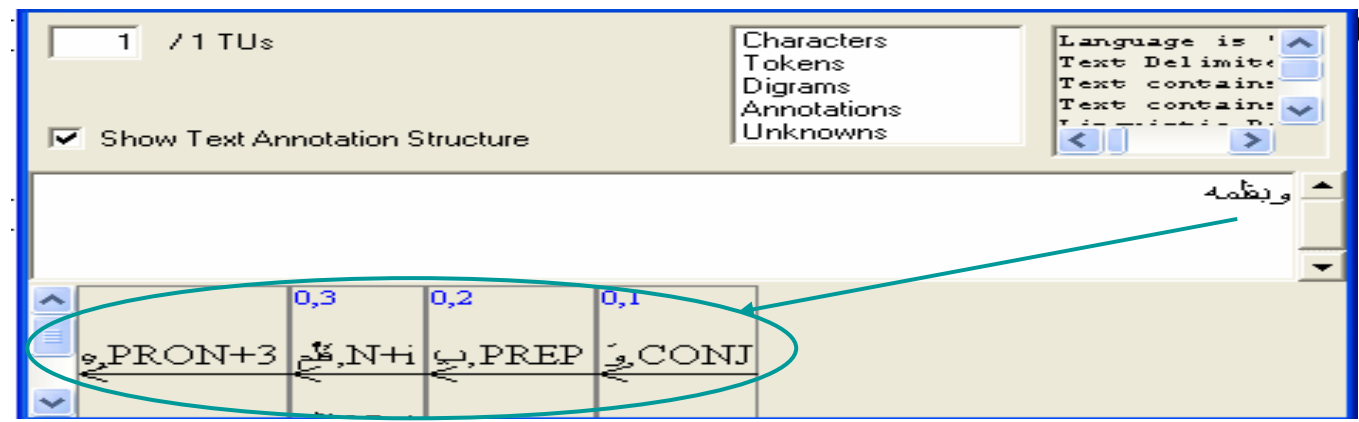
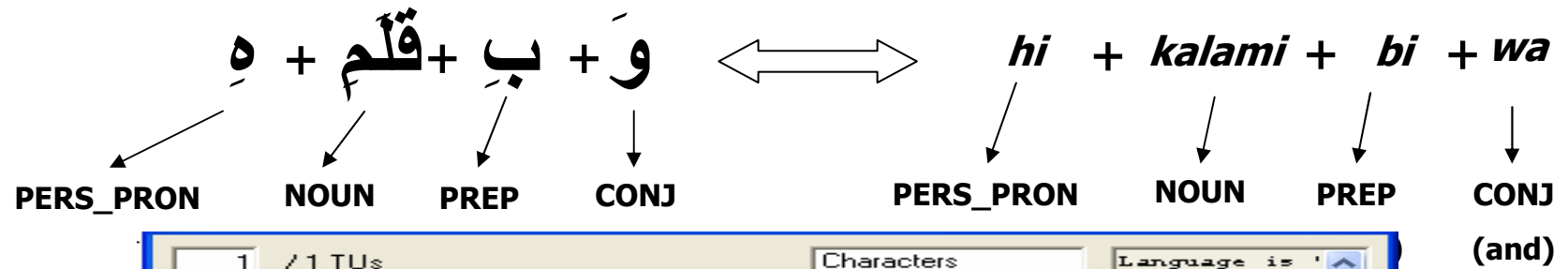
Morphological Analysis (2)

Example:

Text form: «وَبِقَلَمِهِ» (*wabikalamihi* – and with his pen)

1st step: Identification of the radical and affixes

2nd step: Morphological grammars + dictionaries



Lexical constraints

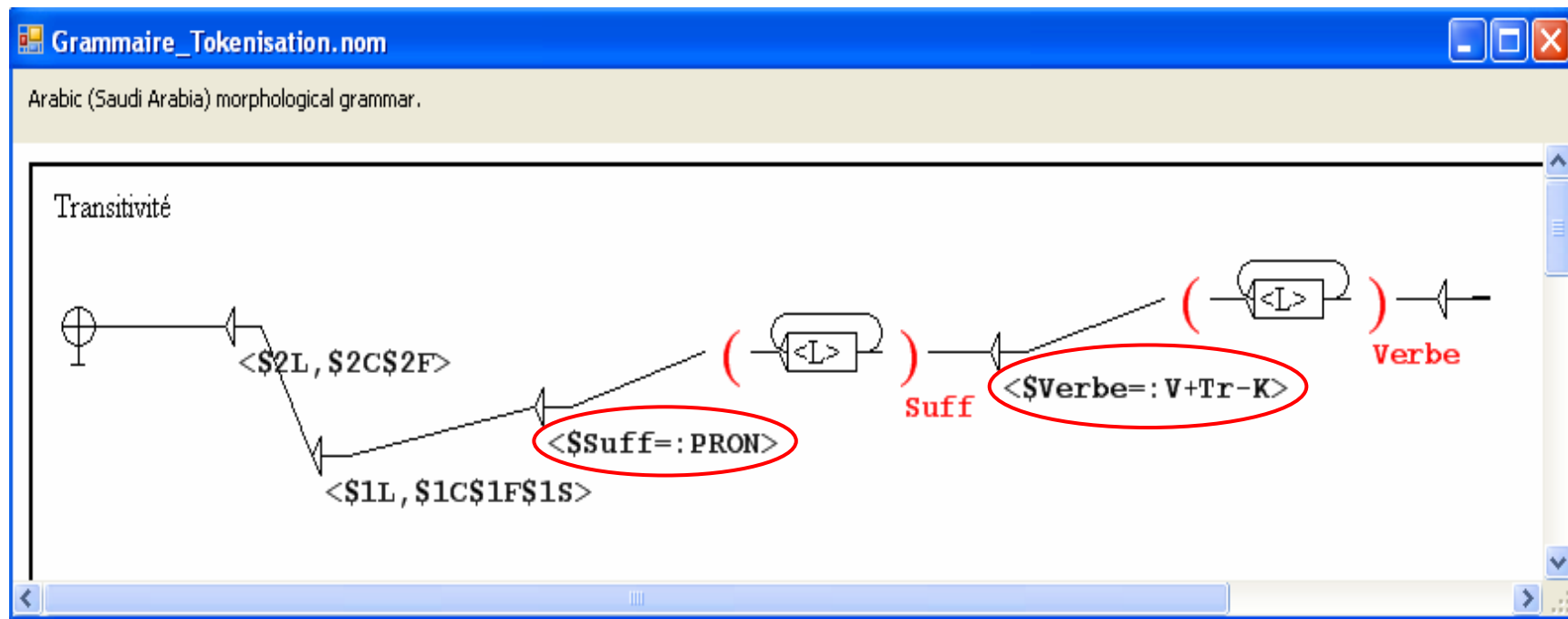
Lexical constraints → working only with valid combinations of the various components of the form

4 types of lexical constraints:

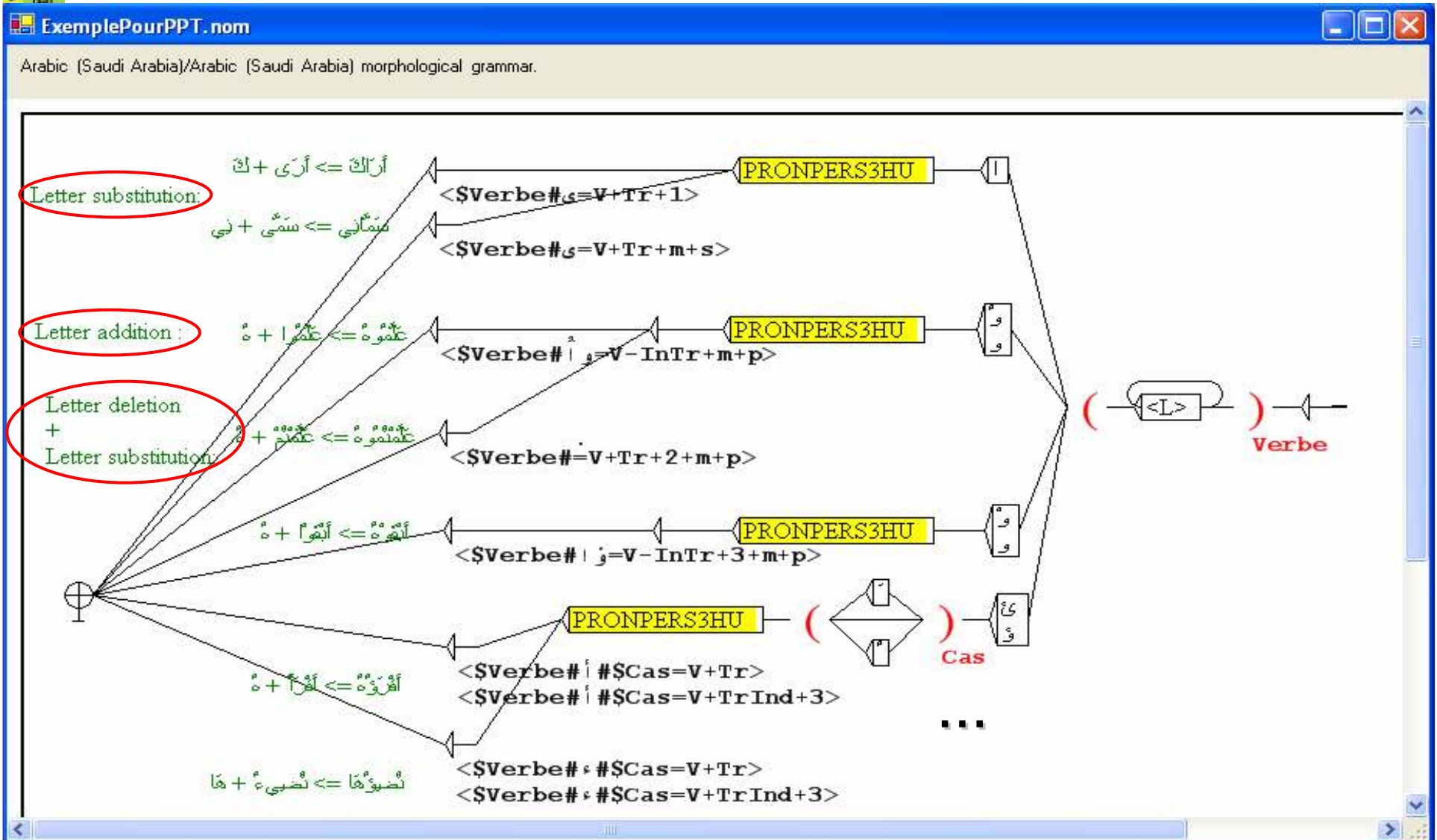
- Constraints on the morpho-syntactic properties of verbs;
- Morphological constraints;
- Orthographical constraints;
- Phonological constraints.

Syntactic constraints on verbs

Transitivity of verbs + non passive voice inflection
 → Possibility of verb suffixation



Morphological constraints



Text annotation

Exemple_Memoire.not

1 / 2 TUs

Characters
Tokens
Digrams
Annotations
Unknowns

Language is "Arabic (Saudi Arabia)(ar)".
Text Delimiter is: \n (NEWLINE)
Text contains 2 Text Units (TUs).
7 tokens including:
7 word forms

Show Text Annotation Structure

هَذَا الزَّيْبِ مَسْوَاةٌ فِي مَسَاوِلَتِ مَجْلِسِ التَّوَابِ

49,01	49	42	30	26	16,01	16	7,01	7	0,0
مَجْلِسِ التَّوَابِ, N+N_PREF_N	مَسْوَاةٌ, N+p+a	فِي, PREP	هَذَا, PRON	الزَّيْبِ, N+m+s+i_Métier	مَسْوَاةٌ, N+a	مَسْوَاةٌ, N+m+s+u_Métier	مَسْوَاةٌ, N+a	مَسْوَاةٌ, N+a	
مَسْوَاةٌ, N+p+i_Métier	مَسْوَاةٌ, N+p+i	مَسْوَاةٌ, N+p+i	مَسْوَاةٌ, N+a	مَسْوَاةٌ, N+m+s+u_Métier	مَسْوَاةٌ, N+a	مَسْوَاةٌ, N+m+s+u_Métier	مَسْوَاةٌ, N+a	مَسْوَاةٌ, N+a	
مَسْوَاةٌ, N+p+u_Métier	مَسْوَاةٌ, N+p+u	مَسْوَاةٌ, N+p+u	مَسْوَاةٌ, N+a	مَسْوَاةٌ, N+m+s+a_Métier	مَسْوَاةٌ, N+a	مَسْوَاةٌ, N+m+s+a_Métier	مَسْوَاةٌ, N+a	مَسْوَاةٌ, N+a	

Application : Spelling checker and corrector

we have implemented finite state cascade transducers to deal with some frequent spelling errors :

- Letter confusion in the beginning of a word form : “ا” (*álif*) vs. “أ” (*hamzä*) ;
- Letter confusion at the end of a word form : “ي” (*yâ'*) vs. “ى” (*álif maqşûrâ*) or “ة” (*h - tâ' marbûţä*) vs. “ه” (*h - hâ'*) ;
- Letter inversion : “ل” vs. “ل” ;

Conclusion

- Build finite state transducers to deal with agglutinative word structure
- Transducers associate words with lexical constraints → work only with valid combinations of the various morphemes of the form
- Segmentations → a higher level of lexical ambiguity → Need to develop disambiguation FSTs

Thank you !