

MT Summit IX Workshop

Machine Translation for Semitic Languages

Semitic Linguistic Phenomena and Variations

Nizar Habash

University of Maryland
Institute for Advanced Computer Studies

Road Map

- Introduction
- Orthography
- Morphology
- Syntax
- Translation Divergences
- Conclusion

Introduction

- What this talk is about
 - Similarities that define “the Semitic family”
 - Variations differentiating members within the family
 - Similarities do not go beyond morphology and syntax
 - Relevance to NLP and MT
- Most researchers focus on one Semitic language
 - Modern Standard Arabic (henceforth, A)
 - Modern Hebrew (henceforth, H)
 - Arabic Dialect: Palestinian Arabic (henceforth, P)

Road Map

- Introduction
- Orthography
 - Phonology
 - Scripts
 - Spelling
 - Ambiguity
- Morphology
- Syntax
- Translation Divergences
- Conclusion

Orthography: *Phonology*

| | Consonant | Vowels | | Script |
|---|-----------|--------|-------------------|-----------------------|
| A | 28 | 6 | 3 short 3 long | יברע 36 graphemes |
| H | 18 | 5 | | תירבע 22 graphemes |
| P | 28 | 10 | 5 short 5 long | יברע? 36 graphemes |

Orthography: *Script*

- Alphabets
 - Graphemic Variants
 - ك ك ك ك (27 out of 36), ڤ ڤ (5 out of 22)
 - Encoding issues
- Optional diacritics
 - Some Vowels سُ سَ سِ سِ سِ
 - Lack of vowel سُ سِ
 - Consonantal Doubling سُ سِ

Orthography: *Spelling*

- Mostly consonantal Spelling
 - ماس = slam = salām, שולש = □lvm = □alom
 - Dual use of (ا w/v j يوا) as consonant and vowel
- Diacritics as semantic markers
 - זָכַר (zaxar male) זָכַר (zaxar to remember)
 - كَتَبَ (kataba to write) كُتِبَ (kutiba to be written)

Orthography: *Spelling*

- Hebrew
 - Full Spelling, “Defective” Spelling (רסה ביתכ, אלמ ביתכ)
 - *kotel* לתכ לתוכ (wall)
- Arabic
 - Morphophonemic Spelling
 - Feminine Marker ة (ta marbuta)
 - *kabīr* big ♂ ريبك (♂) *kabīra* big ♀ ريبك (♀)
 - Derivation Marker
 - *hawa* (to love) (اوہ) (air) (یوہ)
 - Hamza Variants (6 characters for one phoneme)
 - هـ اءب مؤ اءب ء اءب (یؤ اءب آ ء)

Orthography: *Ambiguity*

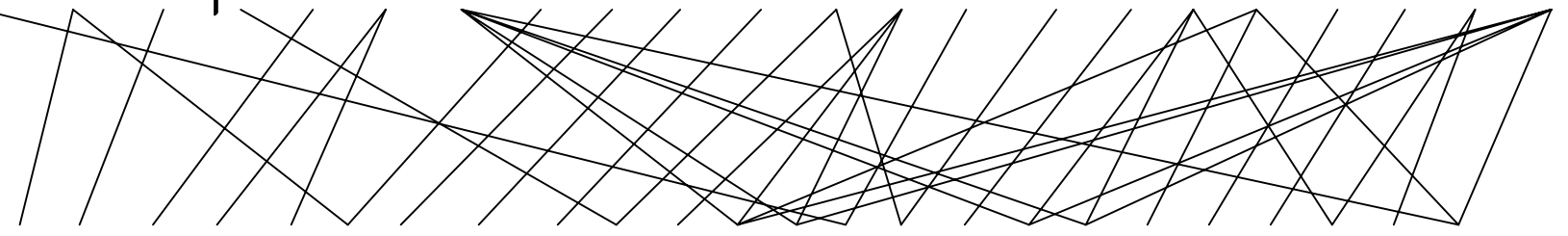
A

ء آ إ وئ ی ا ب ت ة ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي



ī j ū w h n m l k q f □ □ d t ḍ ṣ □ s z r ḏ d x ḥ □ θ t b ā □

א ב ג ד ה ו ז ח ט י כ ל מ נ ס ע פ צ ק ר ש ת

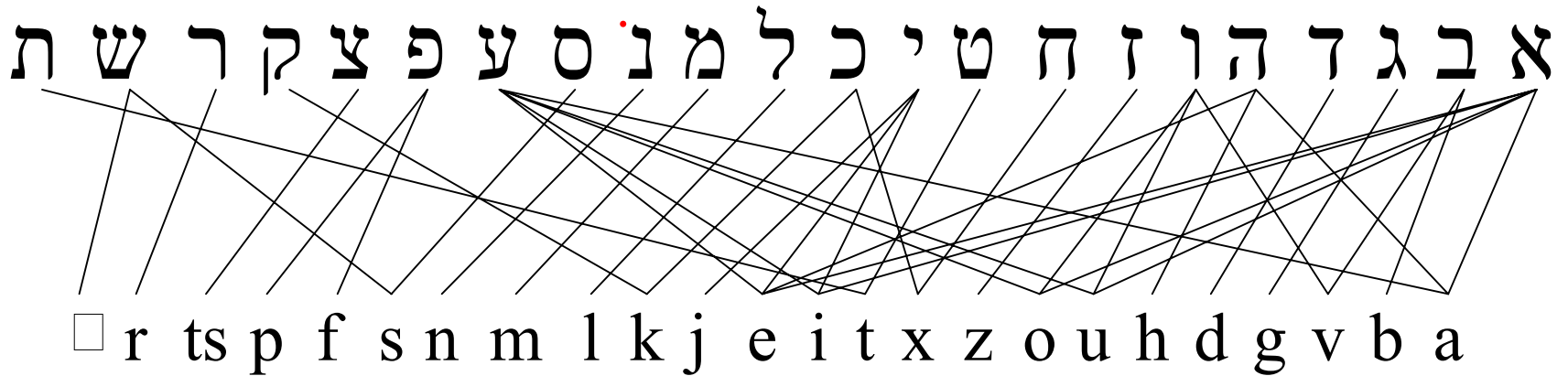
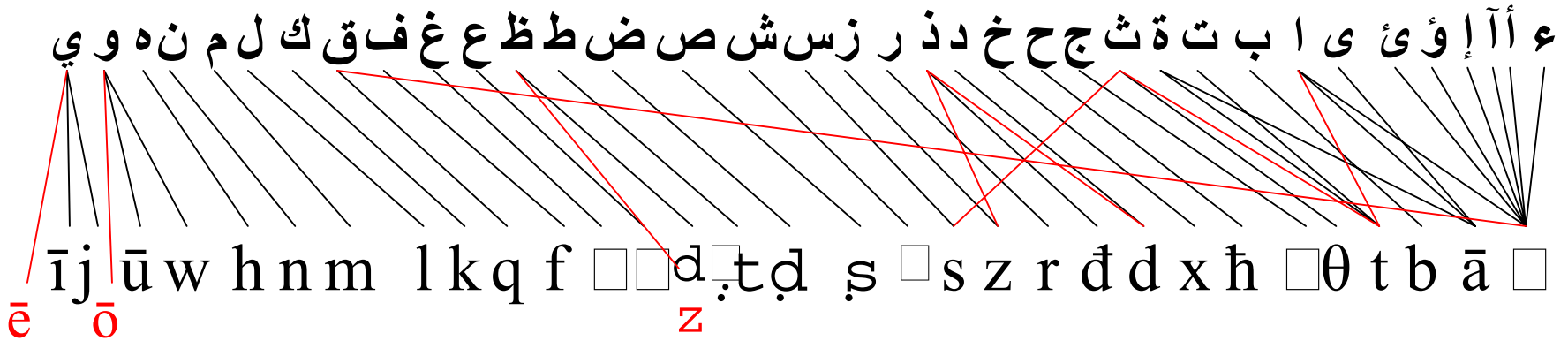


□ r t s p f s n m l k j e i t x z o u h d g v b a

H

Orthography: *Ambiguity*

P



H

Road Map

- Introduction
- Orthography
- Morphology
 - Derivational
 - Inflectional
 - Noun Inflections
 - Verb Inflections
- Syntax
- Translation Divergences
- Conclusion

Morphology: *Derivational*

- Roots and Patterns

ك ت ب
↓ ↓ ↓
? و ? ? م

K T B

כ ת ב
↓ ↓ ↓
? ו ? ?

وتكم

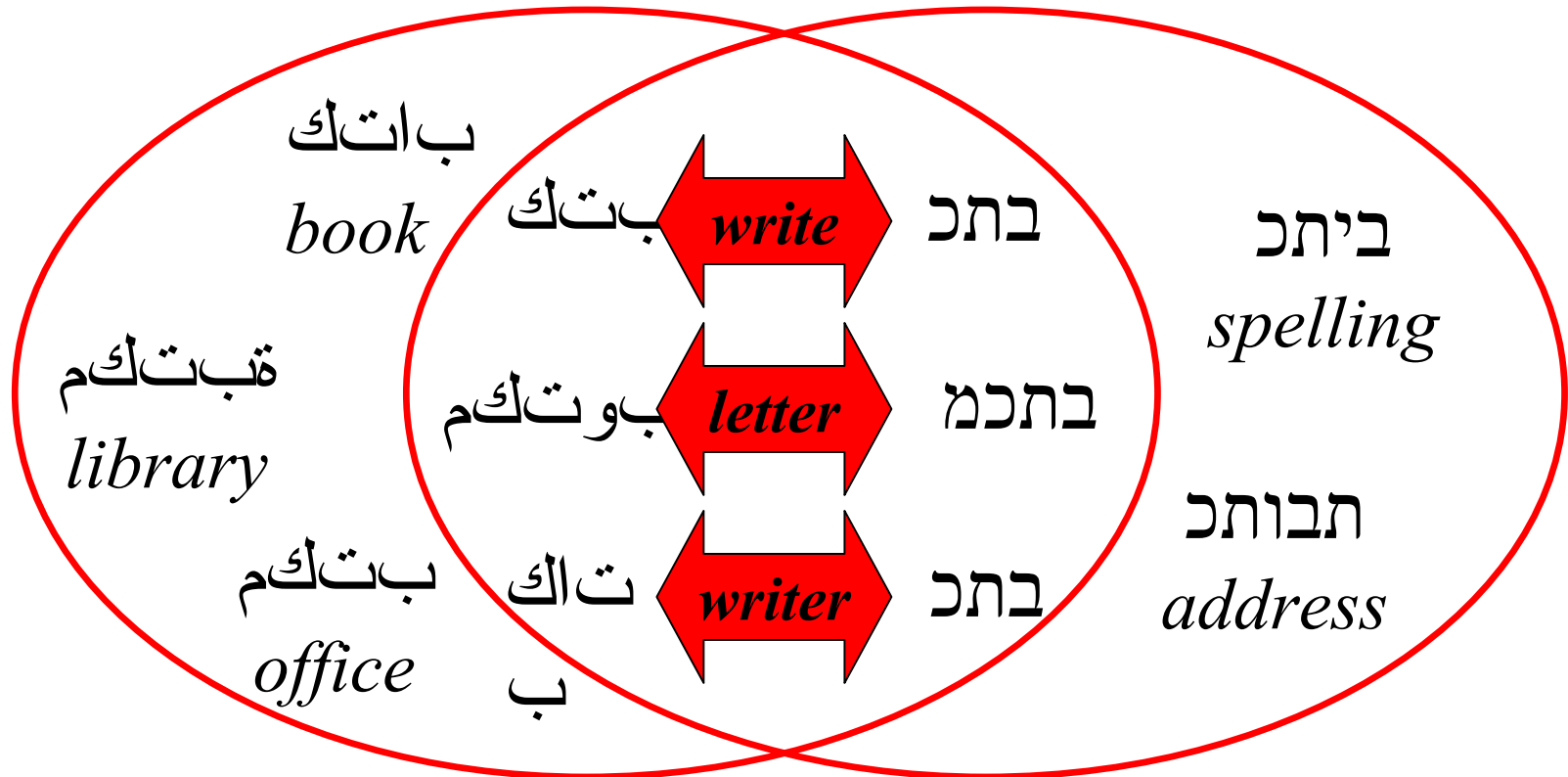
בוּתַכּ

ב
Meaning =

(Root.Meaning+Pattern.Meaning)*Idiosyncrasy.Random

Morphology: *Root Meaning*

- KTB: *writing* “stuff”



Morphology: *Root Meaning*

- LHM-1

מחל
laHm

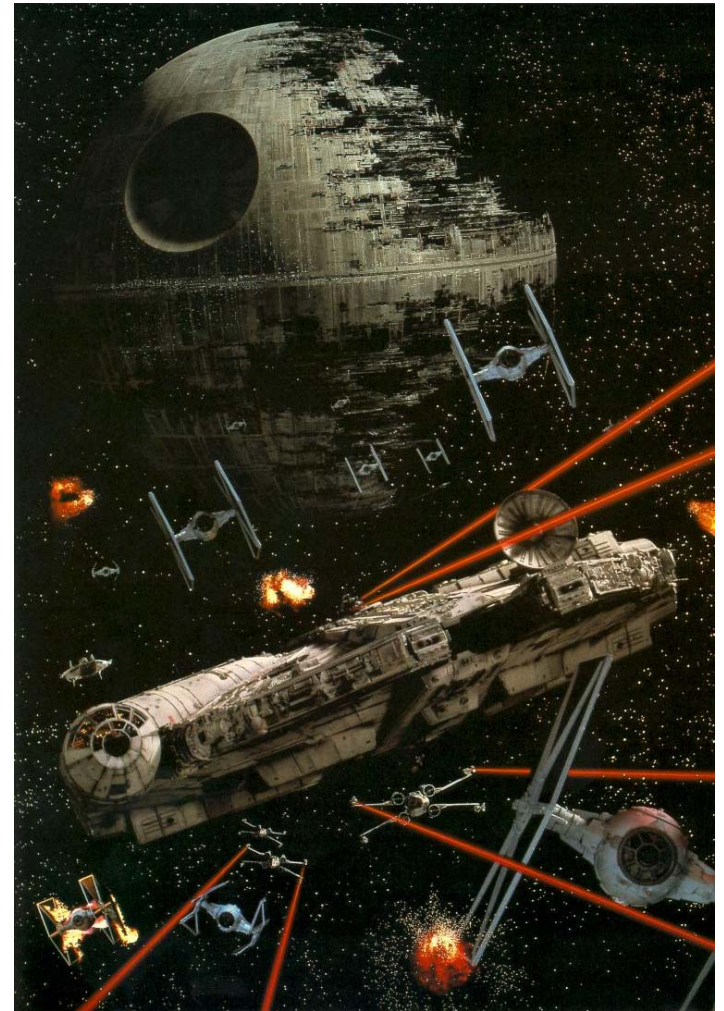


מחל
lexem



Morphology: *Root Meaning*

- LHM-2 (battle sense)
 - قم حل م
 - Fierce battle, massacre, epic
 - המיחל סחול סחל המחול המחלמ
 - War, battle, quarrel, conflict, combat, warfare, belligerence, fighting, quarreling, fighter, militarism, militancy, bellicosity

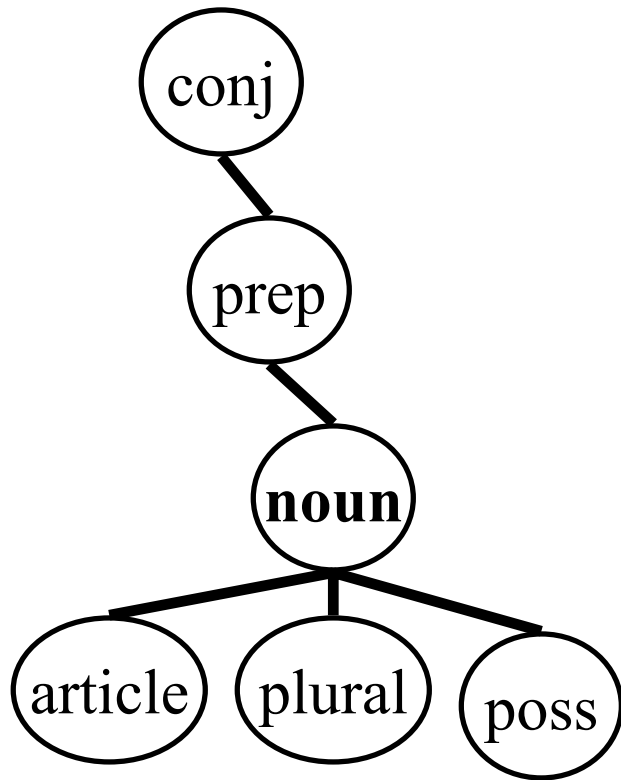


Morphology: *Root Meaning*

- LHM-4 (Conjunctiva sense)
 - ةي م ح ل
 - conjunctiva
 - תי מח ל
 - conjunctiva



Morphology: *Noun Inflections*



ان توي بك و
و + ك + توي ب + ان
And-like-houses-our
And like our houses

תיבבש
תיב+ה+ב+ש
That-in-the-house
Which is in the house

- Arabic Broken Plurals
- Hebrew Ambiguous definiteness

Morphology: *Verb Inflections*

- Perfect Verb Derivation (*Suffixes only*)

| | 1 st Person Singular | 2 nd Person Singular ♂ | 2 nd Person Singular ♀ |
|----------|---------------------------------|-----------------------------------|-----------------------------------|
| A | كُتبتُ katabtu | كُتبتَ katabta | كُتبتِ katabti |
| H | יכתבתי katavti | יכתבתה katavta | יכתבתי katavt |
| P | كُتبت katabt | | كُتبتِ katabti |

- Imperfect Verb Derivation (*Prefix+Suffix*)

| | 1 st Person Singular | 2 nd Person Singular ♂ | 2 nd Person Singular ♀ |
|----------|---------------------------------|-----------------------------------|-----------------------------------|
| A | أكتبُ aktubu | أكتبُ taktubu | أكتبين |
| H | אכתב extov | תכתב textov | תכתבנה taktubāna |
| P | أكتب aktob | أكتب toktob | أكتبين toktobi |

Morphology:

Semantics of Verb Inflections

| | <i>Perfect</i> | <i>Imperfect</i> | | <i>Participle</i> |
|----------|------------------------------|--|--|---|
| H | בתכ <i>katav</i> Past | בותכי <i>jixtov</i> Future | | בתוכ <i>kotev</i> Present |
| A | בתא <i>kataba</i> Past | בתאי <i>jaktubu</i> Present | בתאיס <i>sajaktubu</i> Future | בתא <i>kātib</i> 0-Tense |
| P | בתא <i>katab</i> Past | בתאי <i>jiktob</i> 0-Tense | בתאיח <i>ħajiktob</i> Future | בתאיב <i>kāteb</i> 0-Tense |

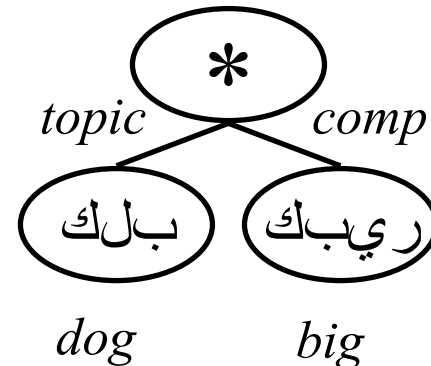
bioktob

Road Map

- Introduction
- Orthography
- Morphology
- **Syntax**
 - **Sentence Structure**
 - **Noun Phrase Structure**
- Translation Divergences
- Conclusion

Sentence Structure

- Sentence structure
 - Copular sentences
 - Verbal sentences
- Copular sentences
 - Topic → Complement
 - Definite → Indefinite
 - **ل** **ري** **بك** **بل** **ك** **ا** **ل** **و** **د** **گ** **ب** **ل** **ك** **ه**
 - **The**-dog big



Sentence Structure

- Verbal sentences
 - *The children wrote the poems*
 - A: **Verb** Subject Object
 - راعش ال ا دال وال ا بتك
 - **Wrote** the-children the-poems
 - H, P: Subject **Verb** Object
 - מירישה תא ובתכ מידליה
 - The-children **wrote** *obj* the-poems
 - راعش ال ا ובتك دال وال ا
 - The-children **wrote** the-poems

Noun Phrase

- Noun → Adjective
- Noun-Adjective Agreement
 - number, gender, definiteness

| | | | |
|---------------------------|--|--------------------------------|--|
| a big dog ♂ | a big dog ♀ | big dogs ♂ | the big dog ♂ |
| ريبيك بلك | قبلك | رابك بالك | ريبيك ^ل ا ^ل بلك ^ل |
| לודג בלכ | ה ^ל לודג ^ל ה ^ל בלכ ^ל | מילודג מיבלכ | לודג ^ה בלכ ^ה |
| <i>dog</i> ♂ <i>big</i> ♂ | <i>dog</i> ♀ <i>big</i> ♀ | <i>dogs</i> ♂ <i>big</i> ♂ +pl | <i>the-dog</i> ♂ <i>the-big</i> ♂ |

Noun Phrase

- **توكيمس / ةفاض** (idafa/smixut)
- **Noun1 of Noun2** encoded structurally
 - Noun1-indefinite → Noun2-definite
 - **וקדרי קלמ נדרל אל كل م**
 - king Jordan = the king of Jordan / Jordan's king
- **Noun1 Form Change**
 - Feminine (H and P)
 - **וקדרי + הכלמ → תכלמ** Queen of Jordan
 - Plural (A and H)
 - **וקדרי + סיכלמ → יכלמ** Kings of Jordan
- Alternatives (only H and P)
 - **Noun1 <particle> Noun2**
 - **וקבת كل م ل ا نדרل** the-king belonging-to Jordan
 - **וקדרי לש קלמה** the-king that-for Jordan

Road Map

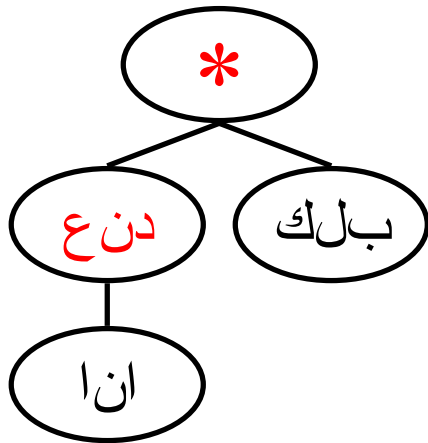
- Introduction
- Orthography
- Morphology
- Syntax
- Translation Divergences
- Conclusion

Translation Divergences

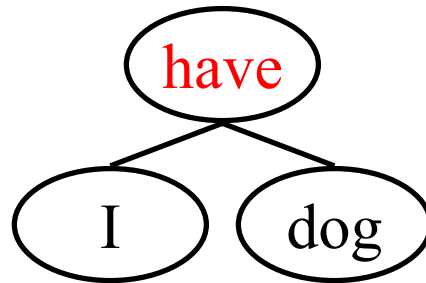
- Variations beyond syntax
- How languages map semantics to syntax
- As complex and diverse as any other language
- Divergence Dimensions
 - Categorical Variation (*develop* → *development*)
 - Conflation (*become frozen* → *freeze*)
 - Inflation (*freeze* → *become frozen*)
 - Structural (*enter the room* → *enter into the room*)
 - Head Swap (*swim across* → *cross swimming*)
 - Thematic (*John likes Mary* → *Mary pleases John*)

Translation Divergences

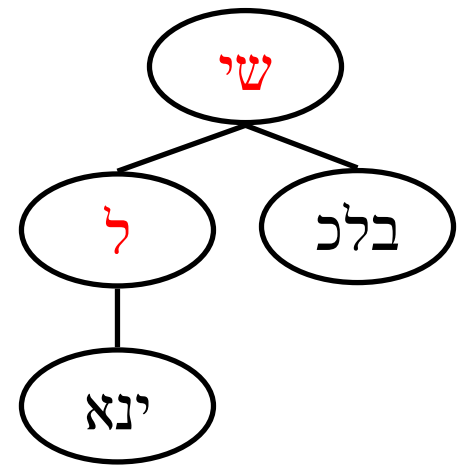
conflation



بلك ي دن ع
at-me dog



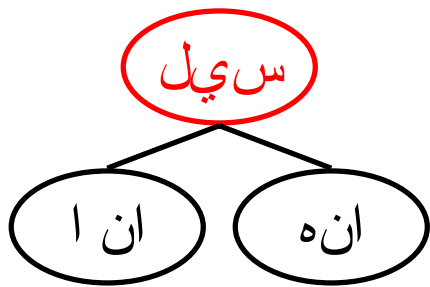
I have a dog



בלק יל שי
there for-me dog

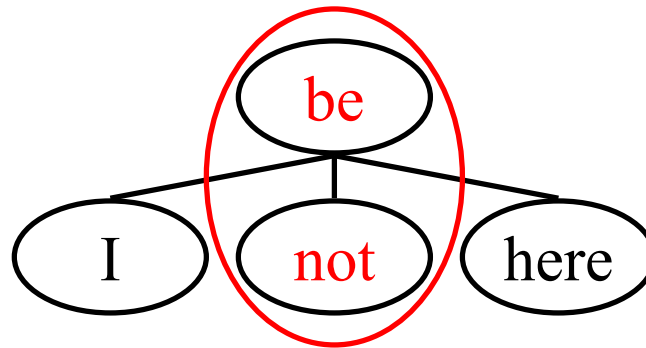
Translation Divergences

conflation

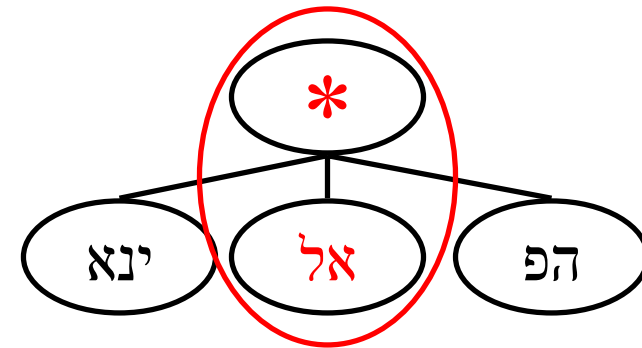


انہ تسئل

I-am-not here



I am not here

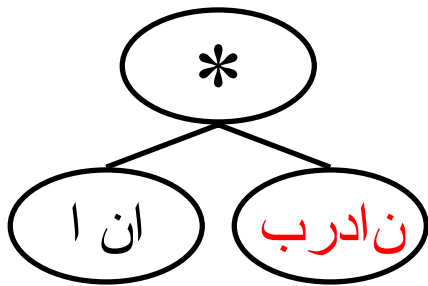


ינא הפ אל

I not here

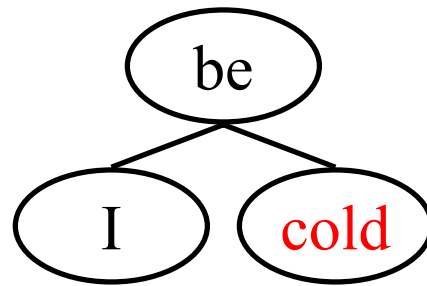
Translation Divergences

thematic

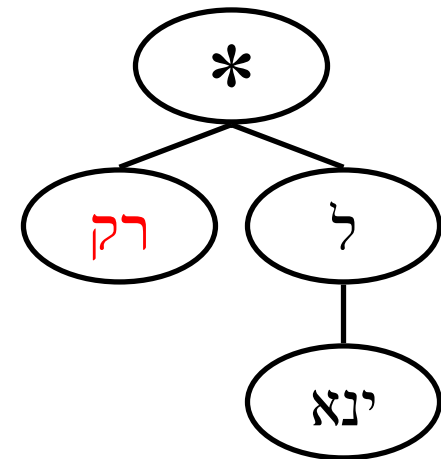


אני קרב

I cold



I am cold

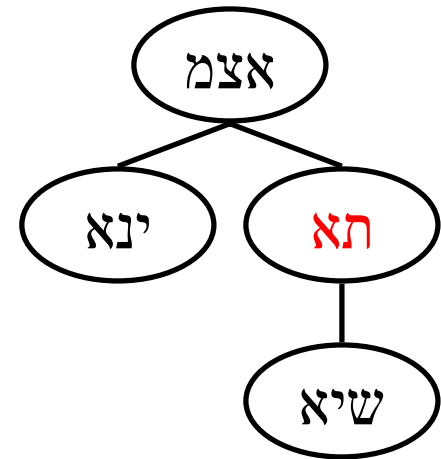
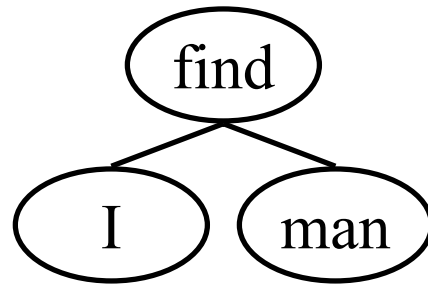
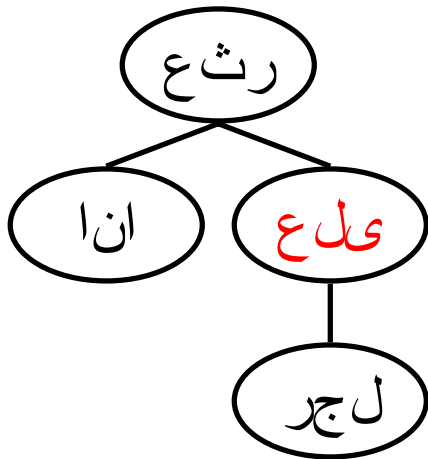


קר לי

cold for-me

Translation Divergences

structural



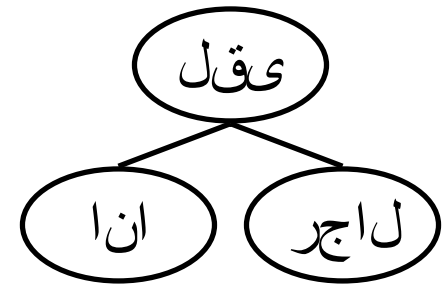
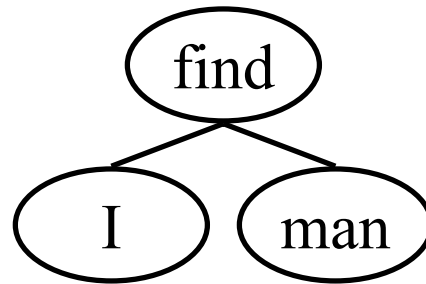
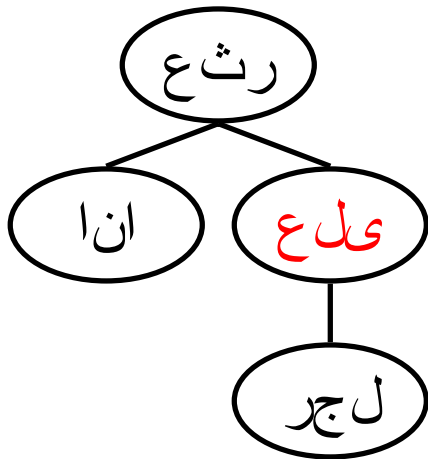
לجرت على انا ترثع
found-I *upon* the-man

I found the man

שיאה תא יתאצמ
found-I *obj* the-man

Translation Divergences

structural



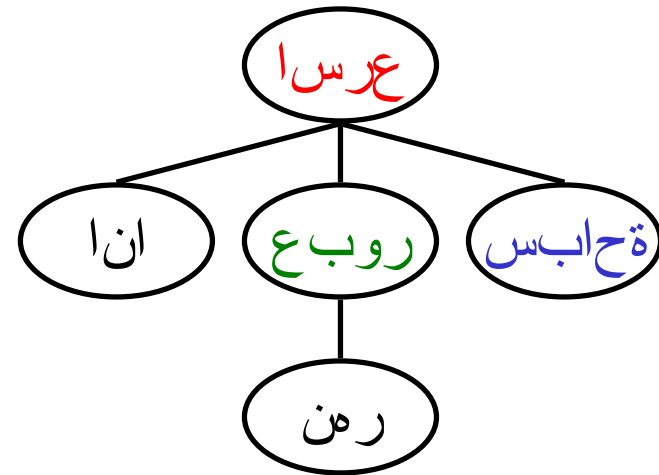
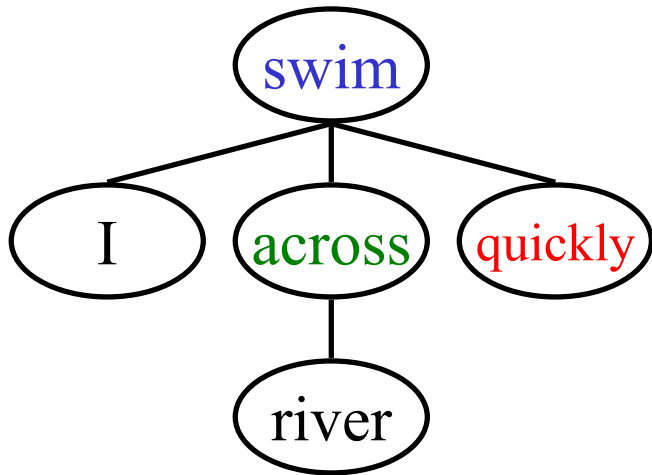
لجر لثقل انا رثع
found-I *upon* the-man

I found the man

لجر لثقل انا
found-I the-man

Translation Divergences

head swap and categorial



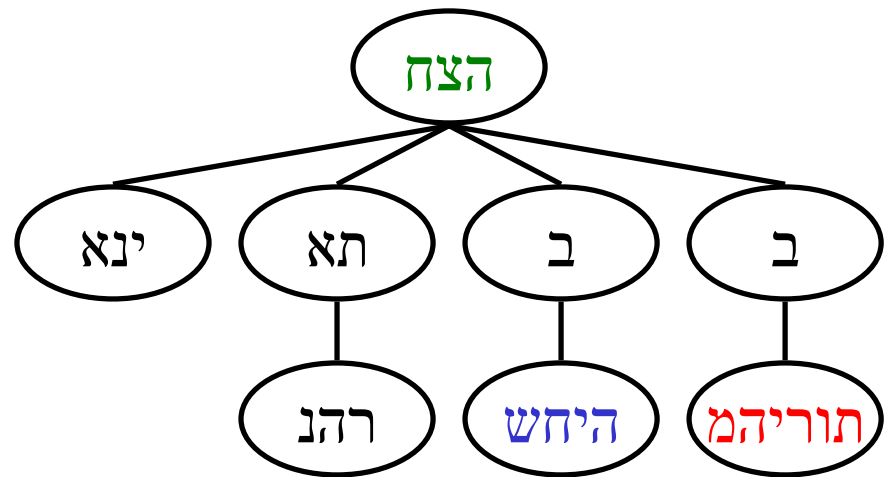
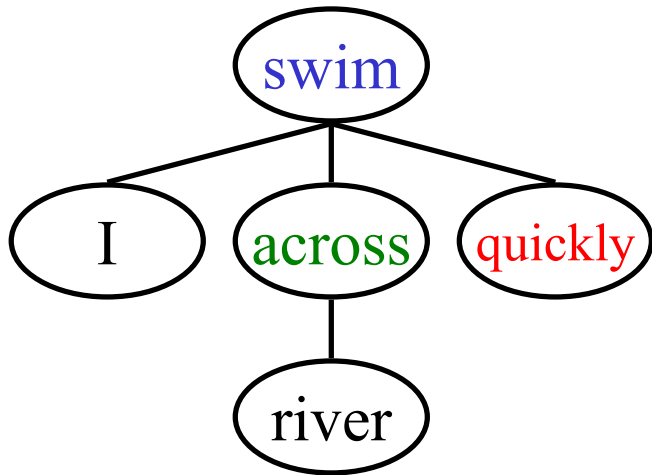
I swam across the river quickly

ةحابس رهن ل ا ر و ب ع ت ع ر س ا

I-sped crossing the-river swimming

Translation Divergences

head swap and categorial



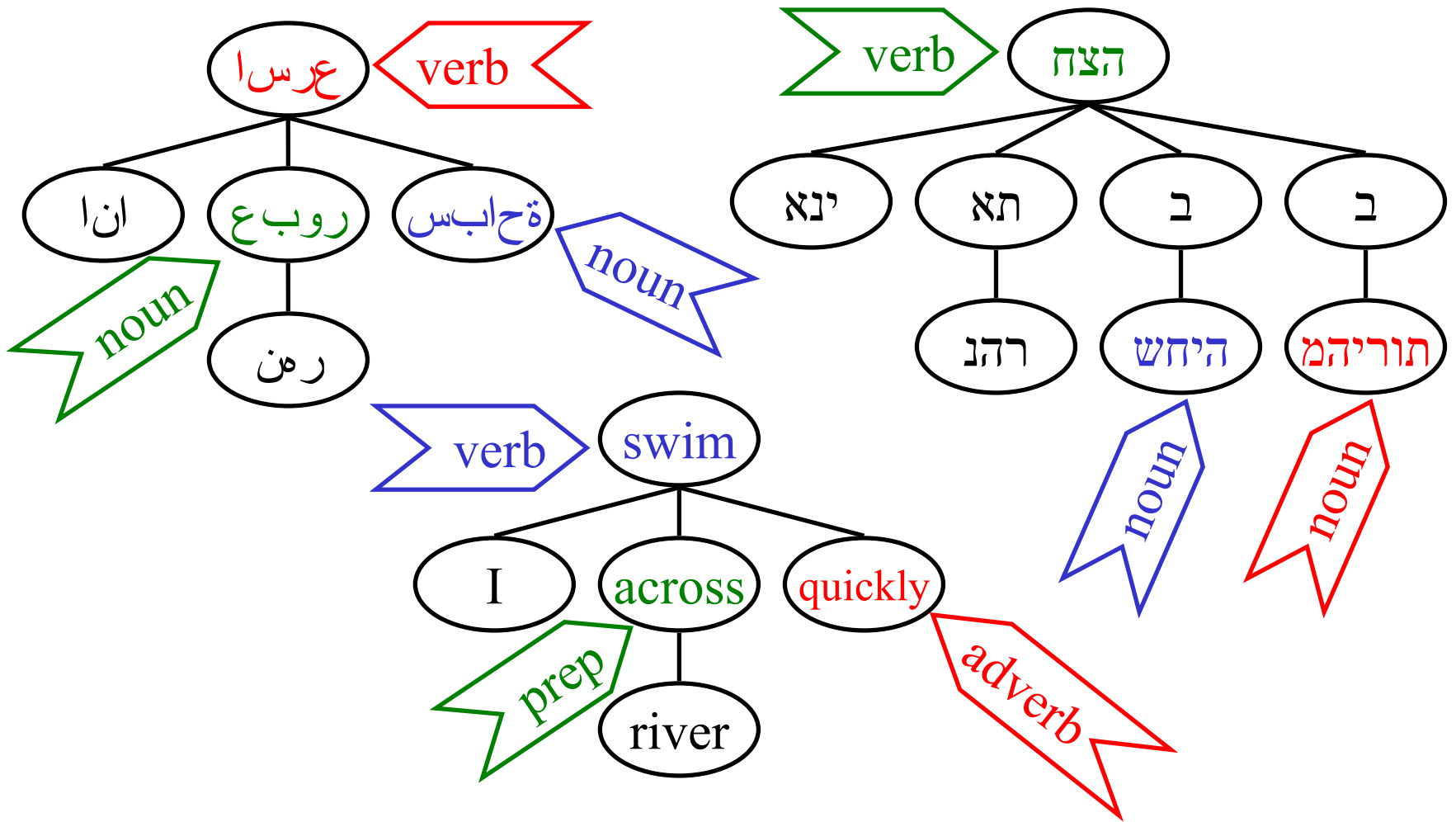
I swam across the river quickly

תוריהםב היחשב רהנה תא יתיצה

I-crossed *obj* river in-swim speedily

Translation Divergences

head swap and categorial



Conclusion

- **Many defining features of the Semitic family**
 - Orthographic conventions, morphological derivation and inflection, phrase structure, etc
- **Many variations that create different kinds of ambiguities and problems**
 - Phonology of orthography, Semantics of derivation and inflection
- **Do similarities extend beyond morphology and syntax?**
 - Translation divergences within Semitic family
 - Ambiguity preservation between Semitic languages