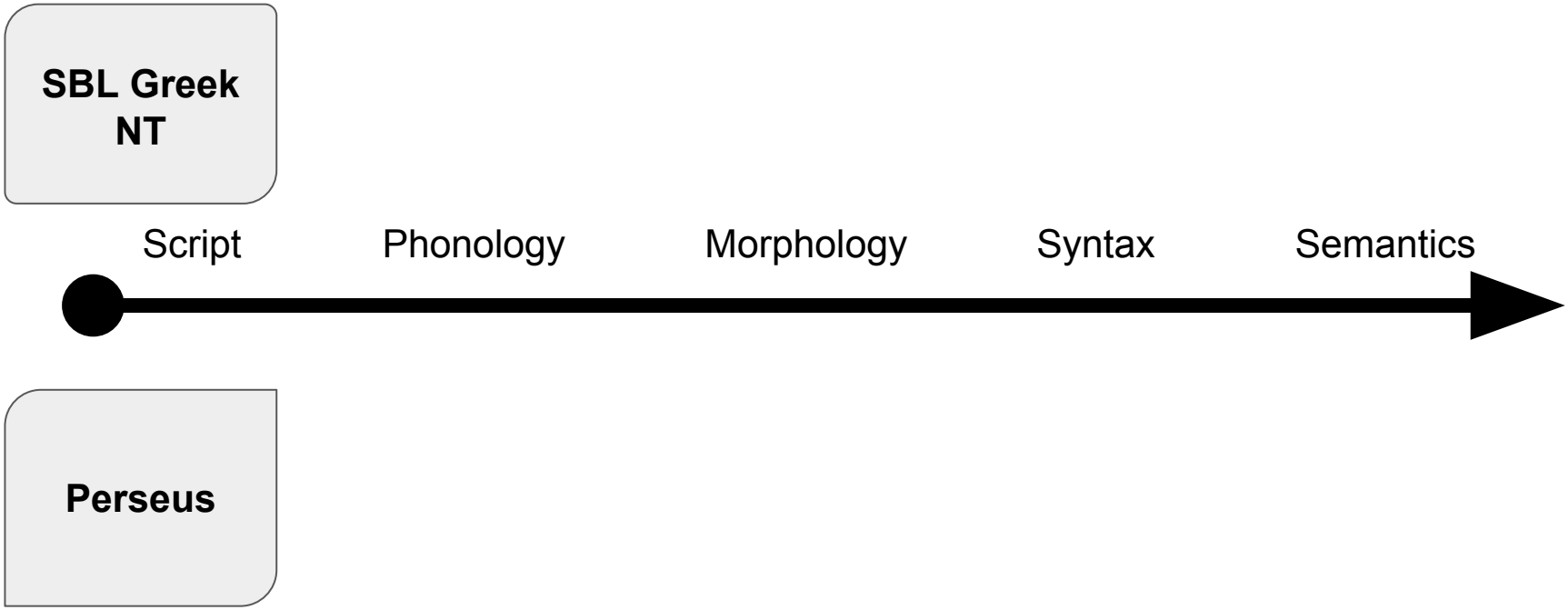
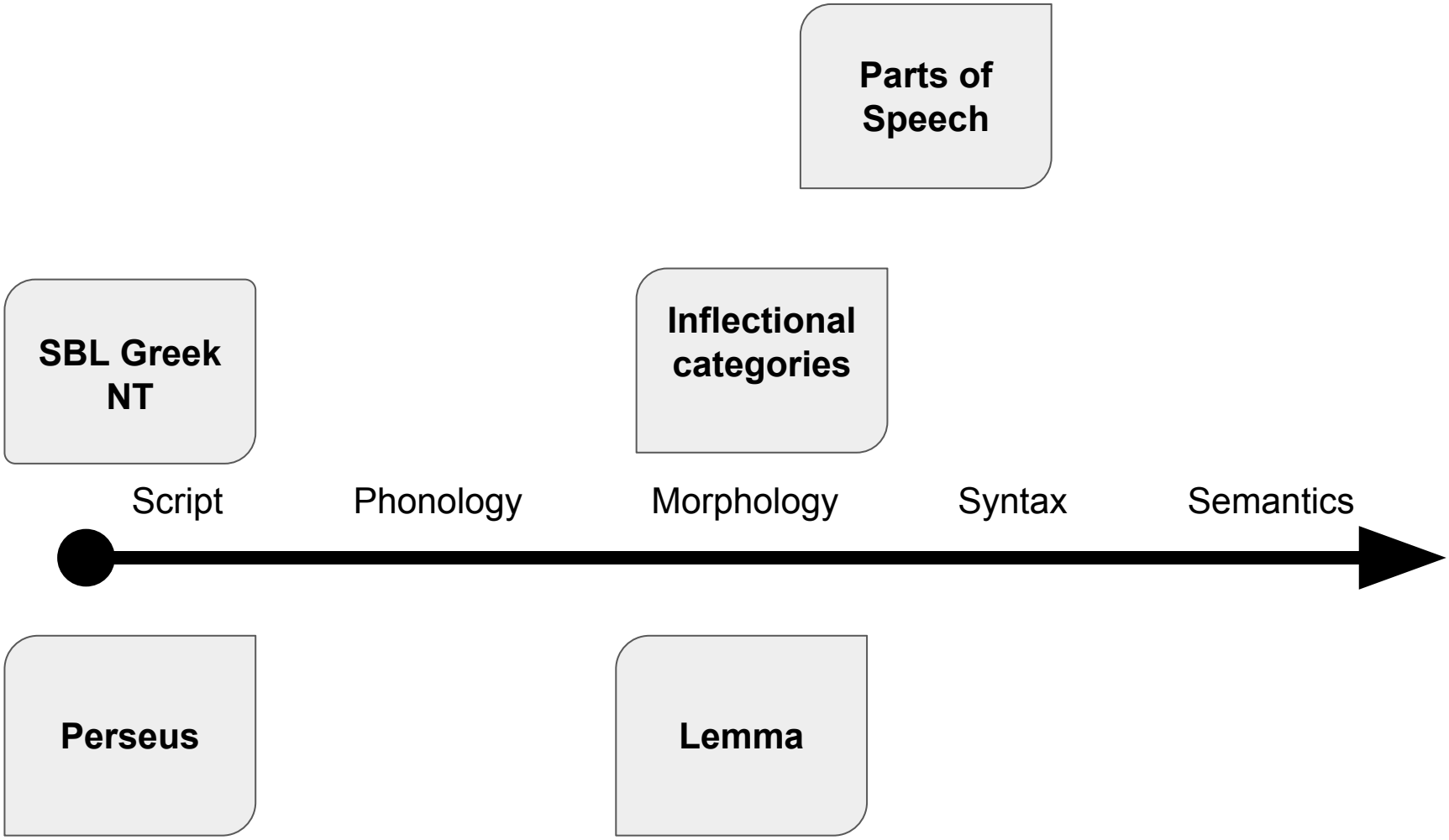
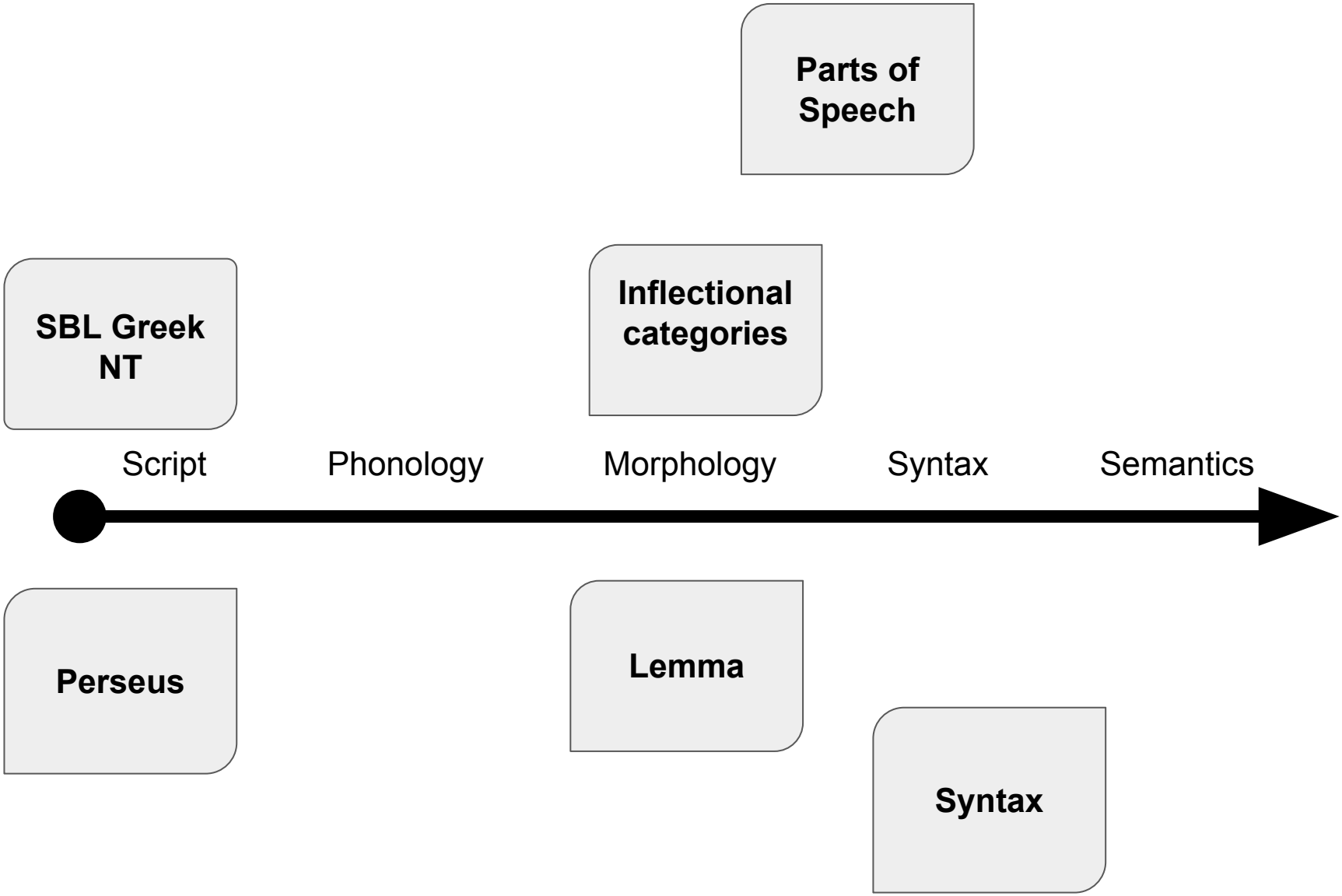
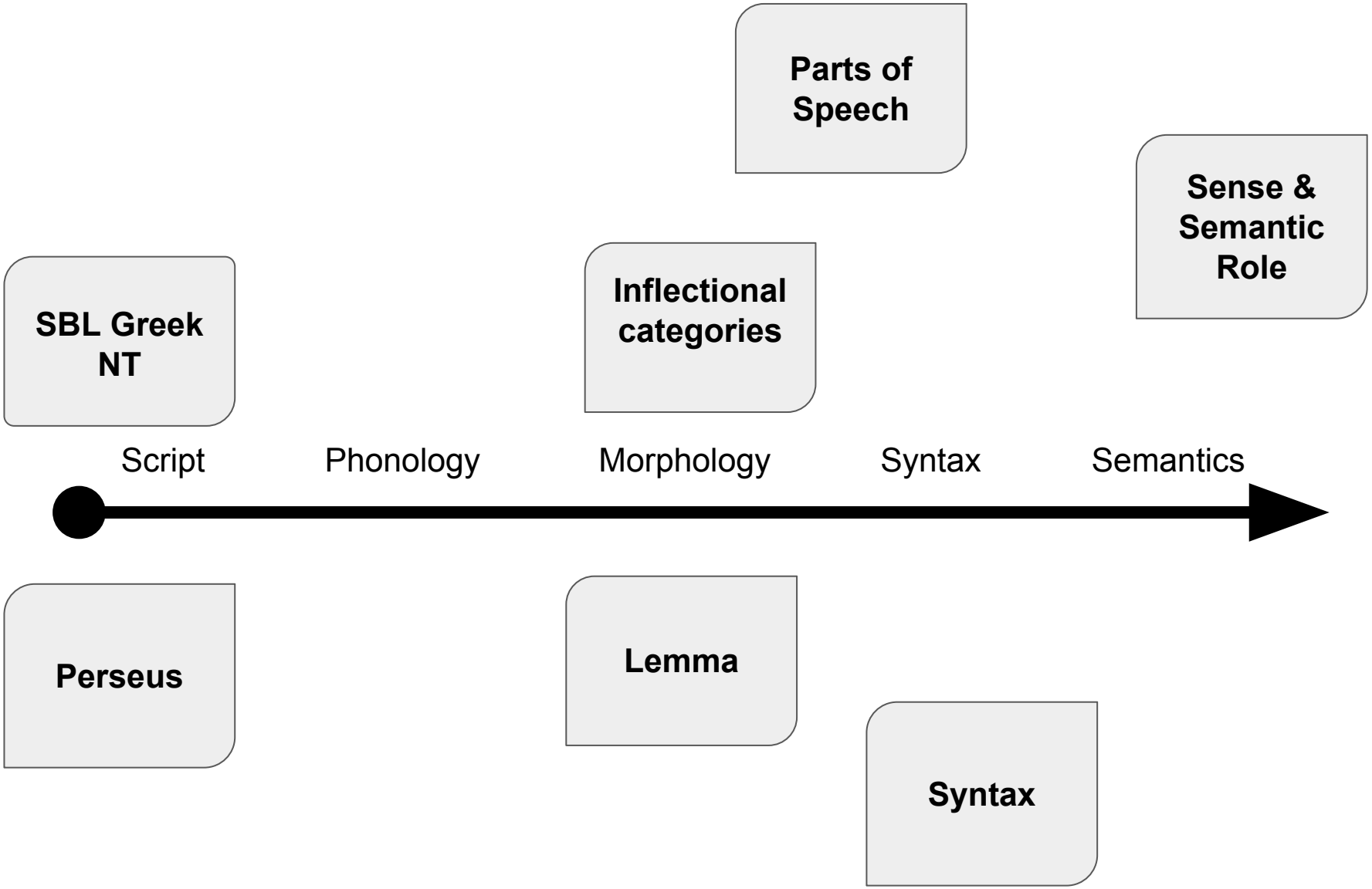
Towards a Comprehensive Computational System for Grammatical Analysis of Ancient Corpora

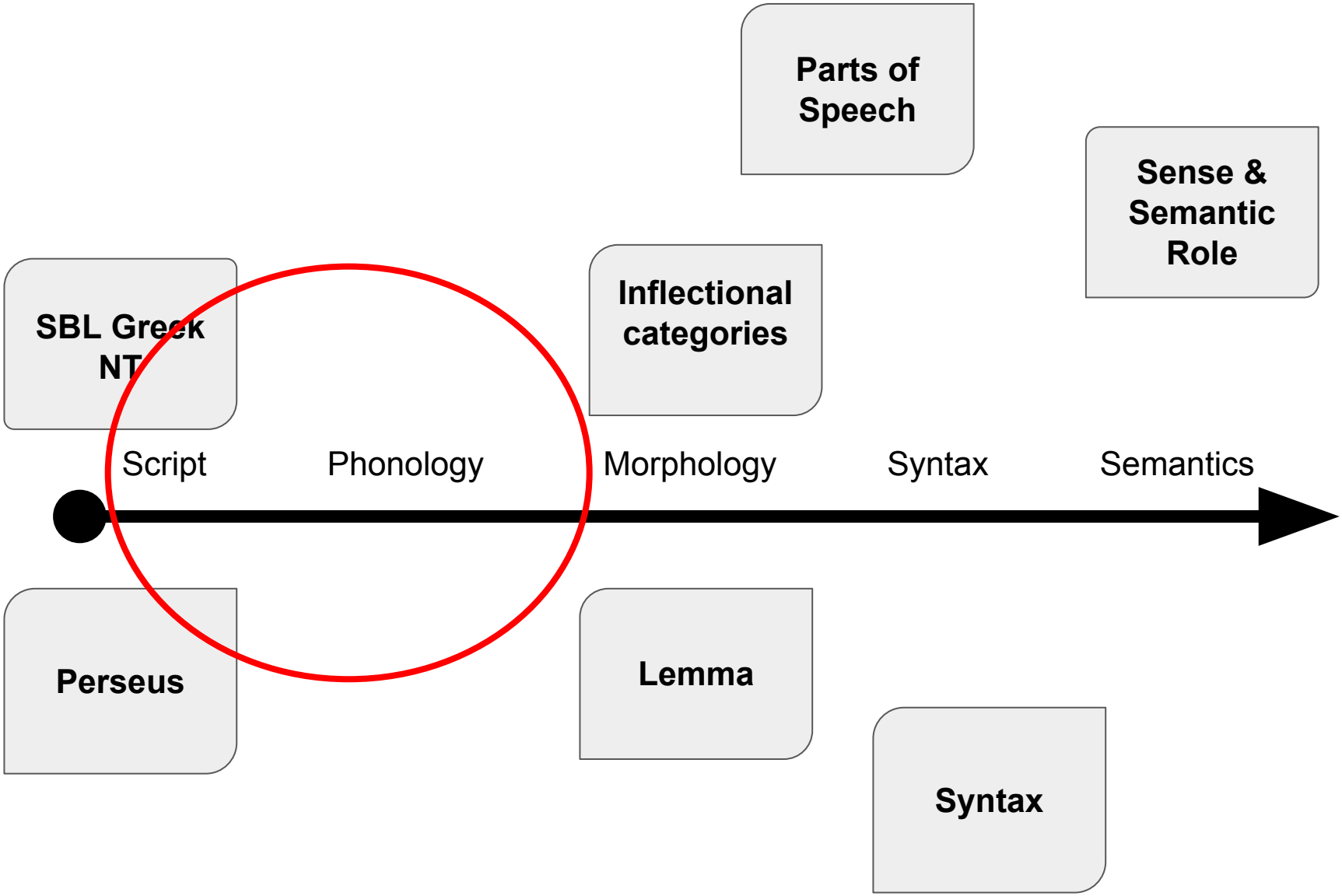
Scott Fleischman and Josh Westbury
Global Education and Research Technology Session

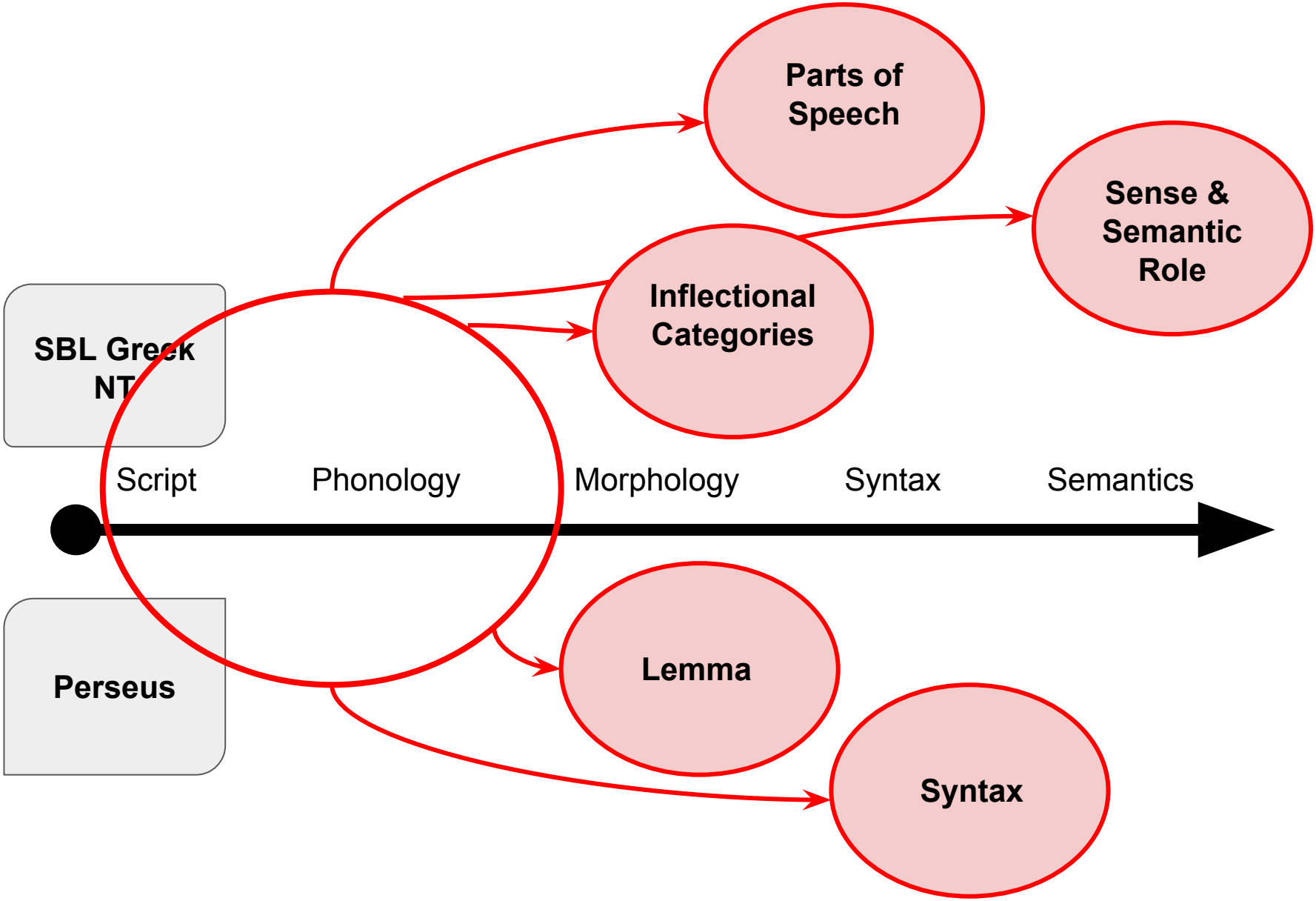

SBL Greek NT
Script
Phonology
Morphology
Syntax
Semantics
Perseus


Parts of Speech
SBL Greek NT
Inflectional categories
Script
Phonology
Morphology
Syntax
Semantics
Perseus
Lemma


Parts of Speech
SBL Greek NT
Inflectional categories
Script
Phonology
Morphology
Syntax
Semantics
Perseus
Lemma
Syntax

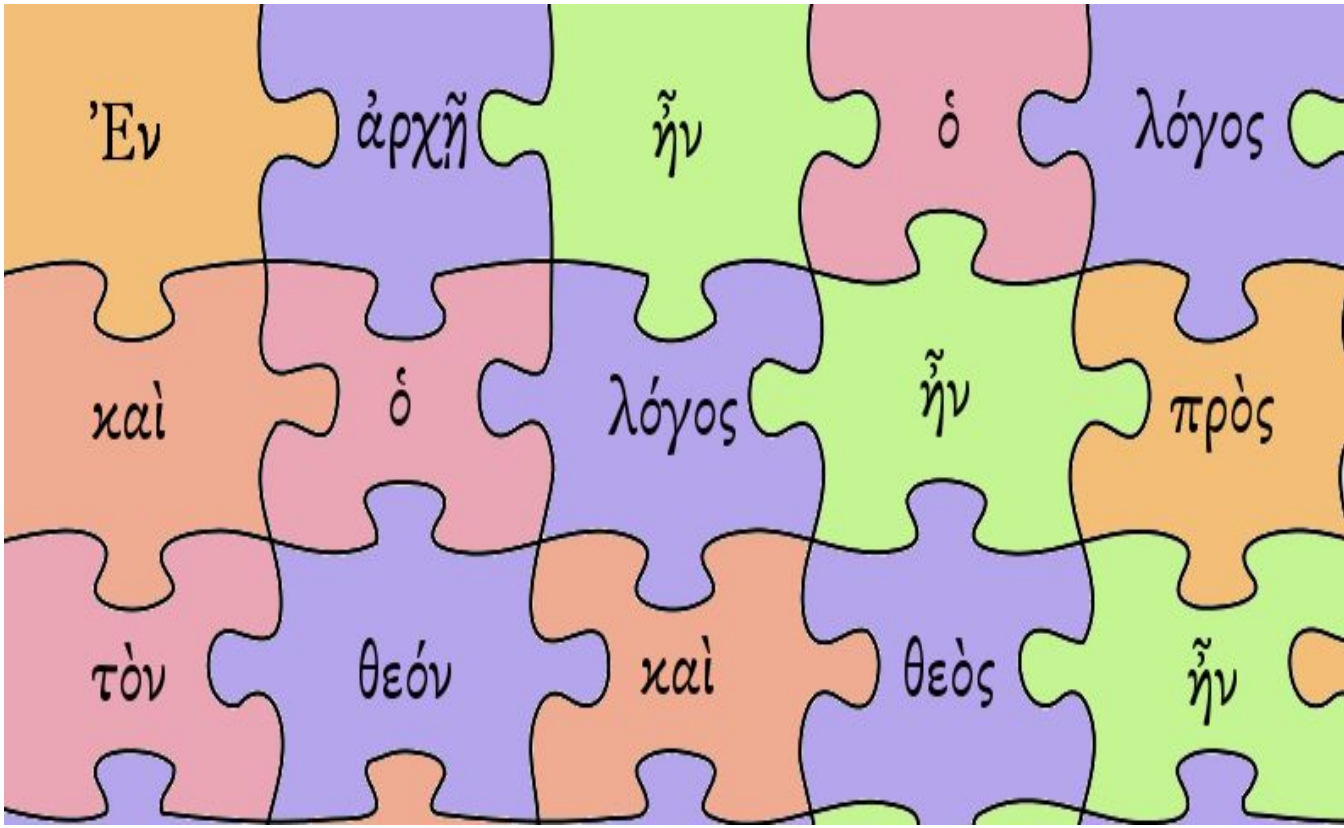

Parts of Speech
Sense & Semantic Role
SBL Greek NT
Inflectional categories
Script
Phonology
Morphology
Syntax
Semantics
Perseus
Lemma
Syntax

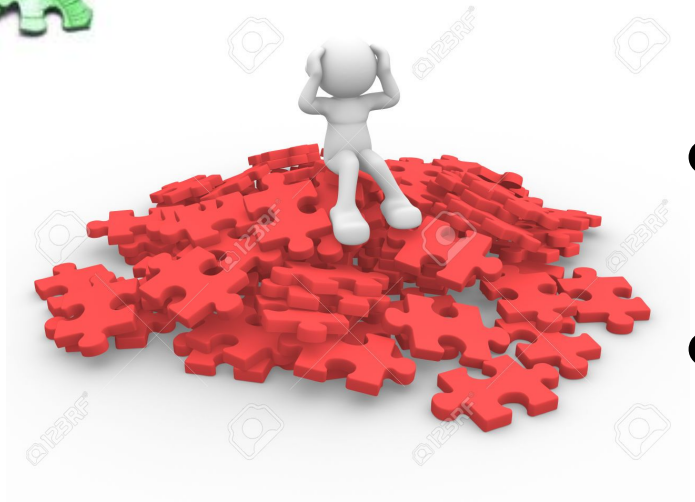

Parts of Speech
Sense & Semantic Role
SBL Greek NT
Inflectional categories
Script
Phonology
Morphology
Syntax
Semantics
Perseus
Lemma
Syntax


Parts of Speech
Sense & Semantic Role
Inflectional Categories
SBL Greek NT
Script
Phonology
Morphology
Syntax
Semantics
Perseus
Lemma
Syntax

Existing Approaches

- Manual
- Time Consuming
- Static
- Isolated

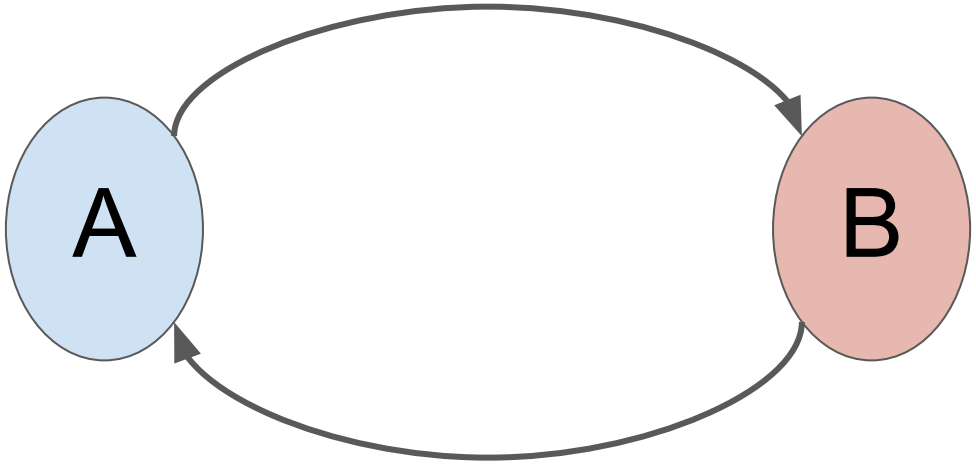




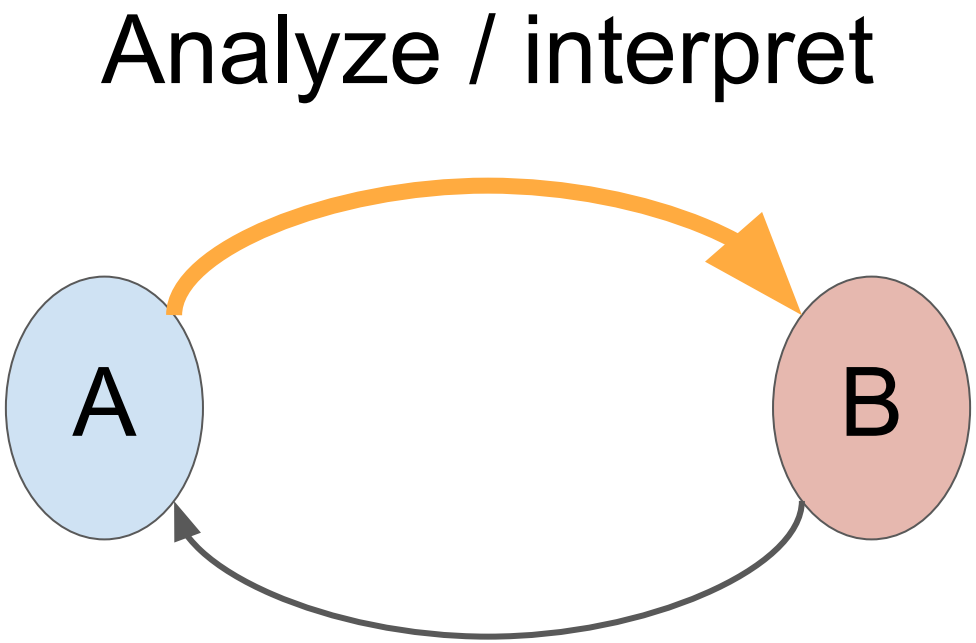
Our Approach

- Automated and Human Directed
- Consistent
- Efficient
- Dynamic
- Unified

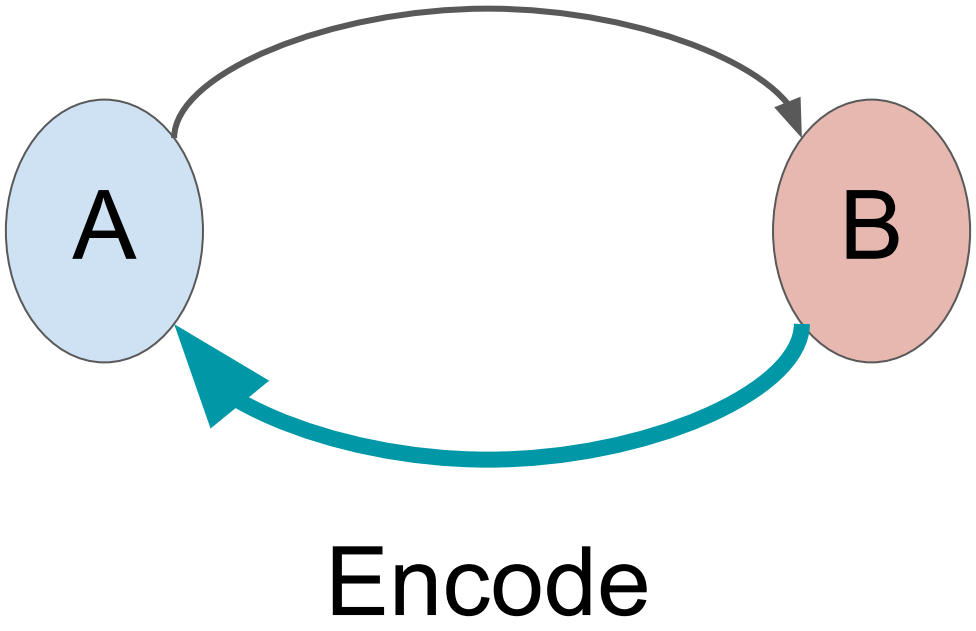
Analysis by transformation



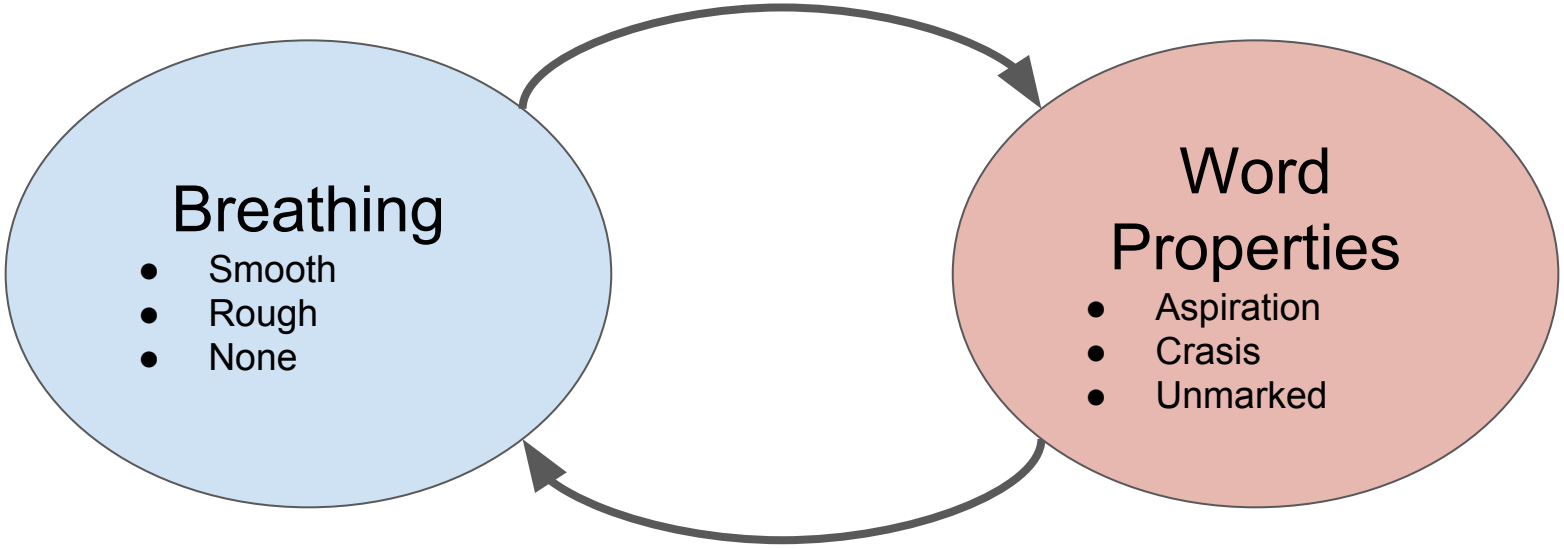
Analysis by transformation



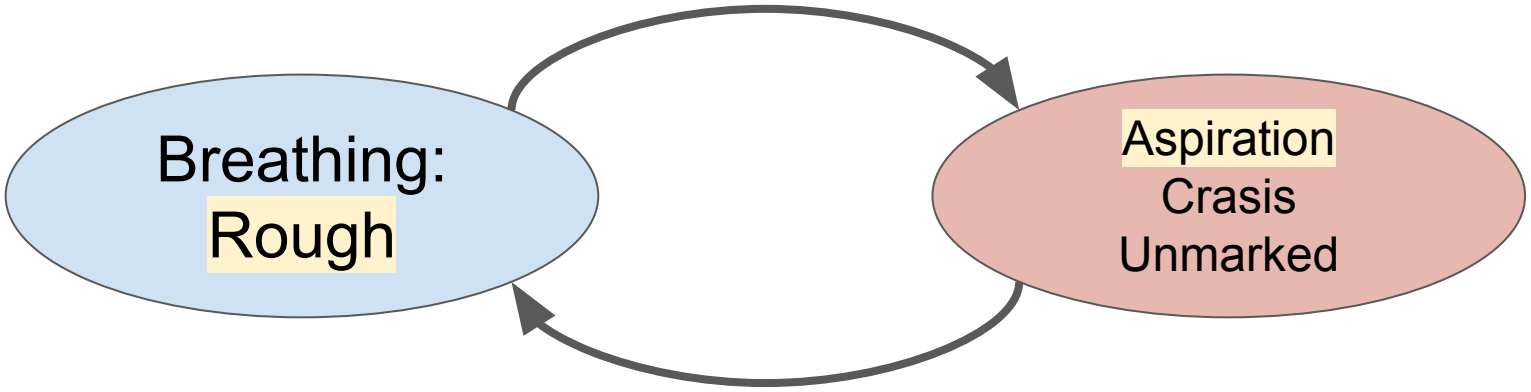
Analysis by transformation



Example: Breathing



Example Summary: Breathing



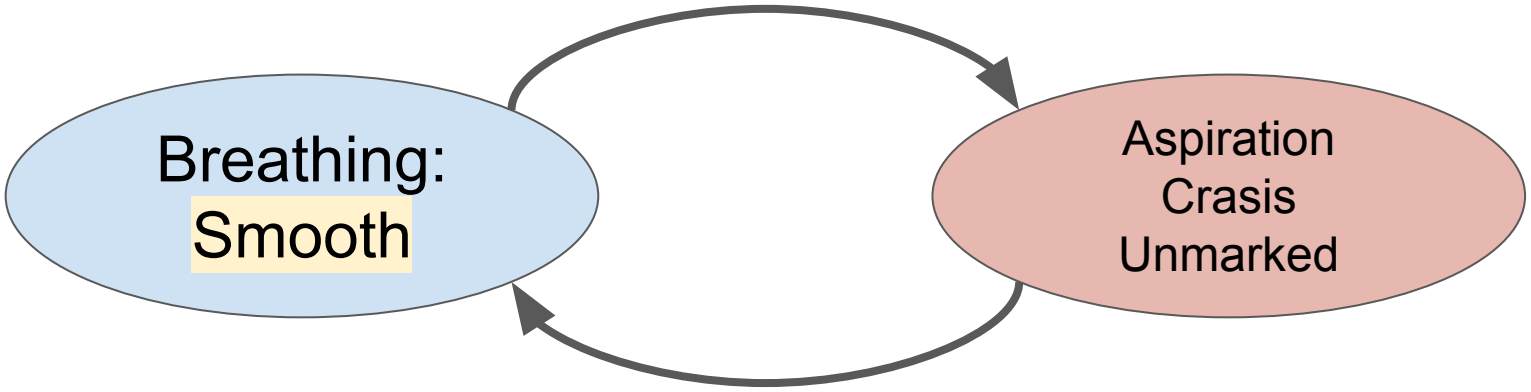
ἡμων ↔ ημων + aspiration

αὐτου ↔ αυτου

κἀγω ↔ καγω + crasis

Αββα ↔ Αββα + unmarked

Example Summary: Breathing



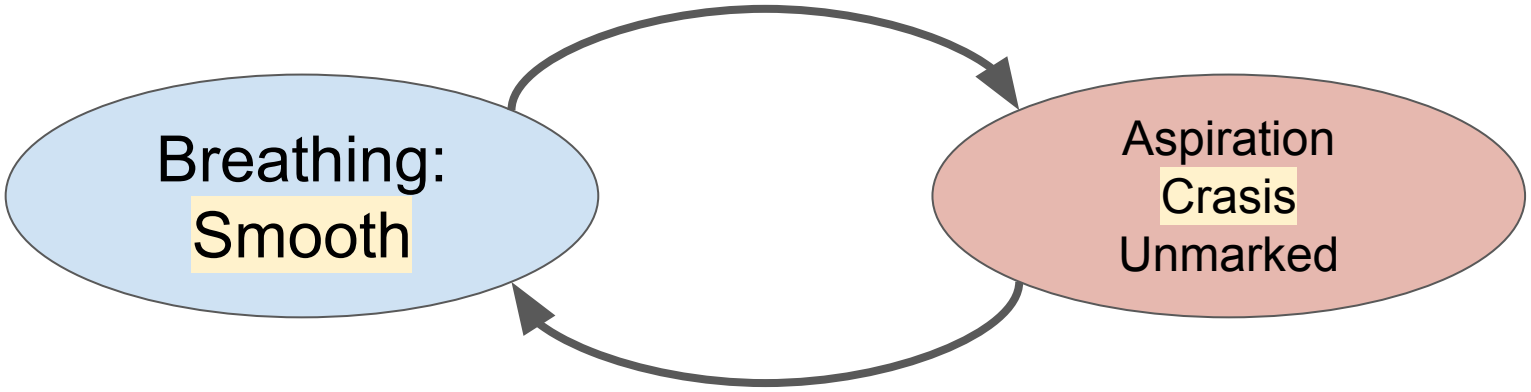
ἡμων ↔ ημων + aspiration

αὐτου ↔ αυτου

κἀγω ↔ καγω + crasis

Αββα ↔ Αββα + unmarked

Example Summary: Breathing



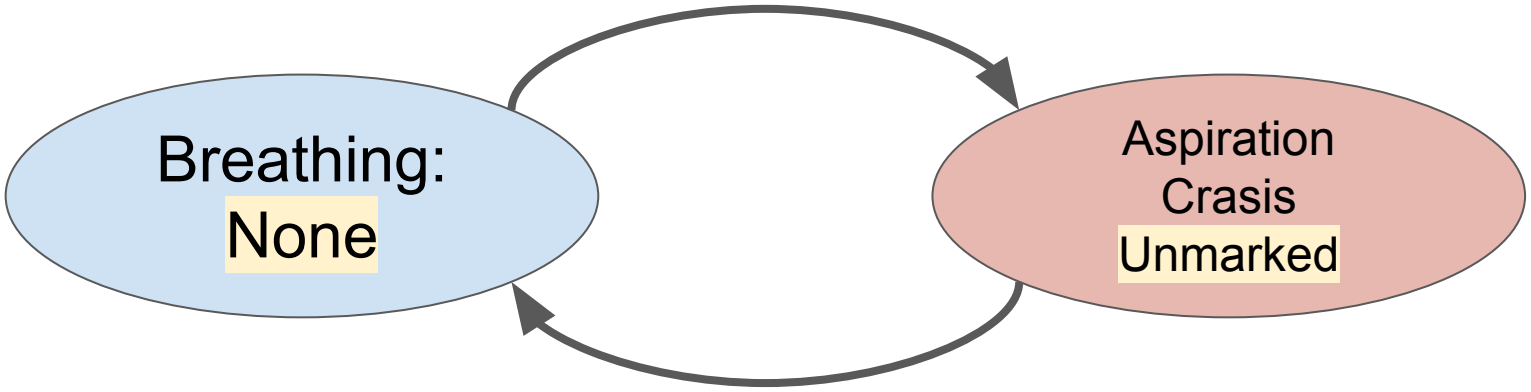
ἡμων ↔ ημων + aspiration

αὐτου ↔ αυτου

κἀγω ↔ καγω + crasis

Αββα ↔ Αββα + unmarked

Example Summary: Breathing



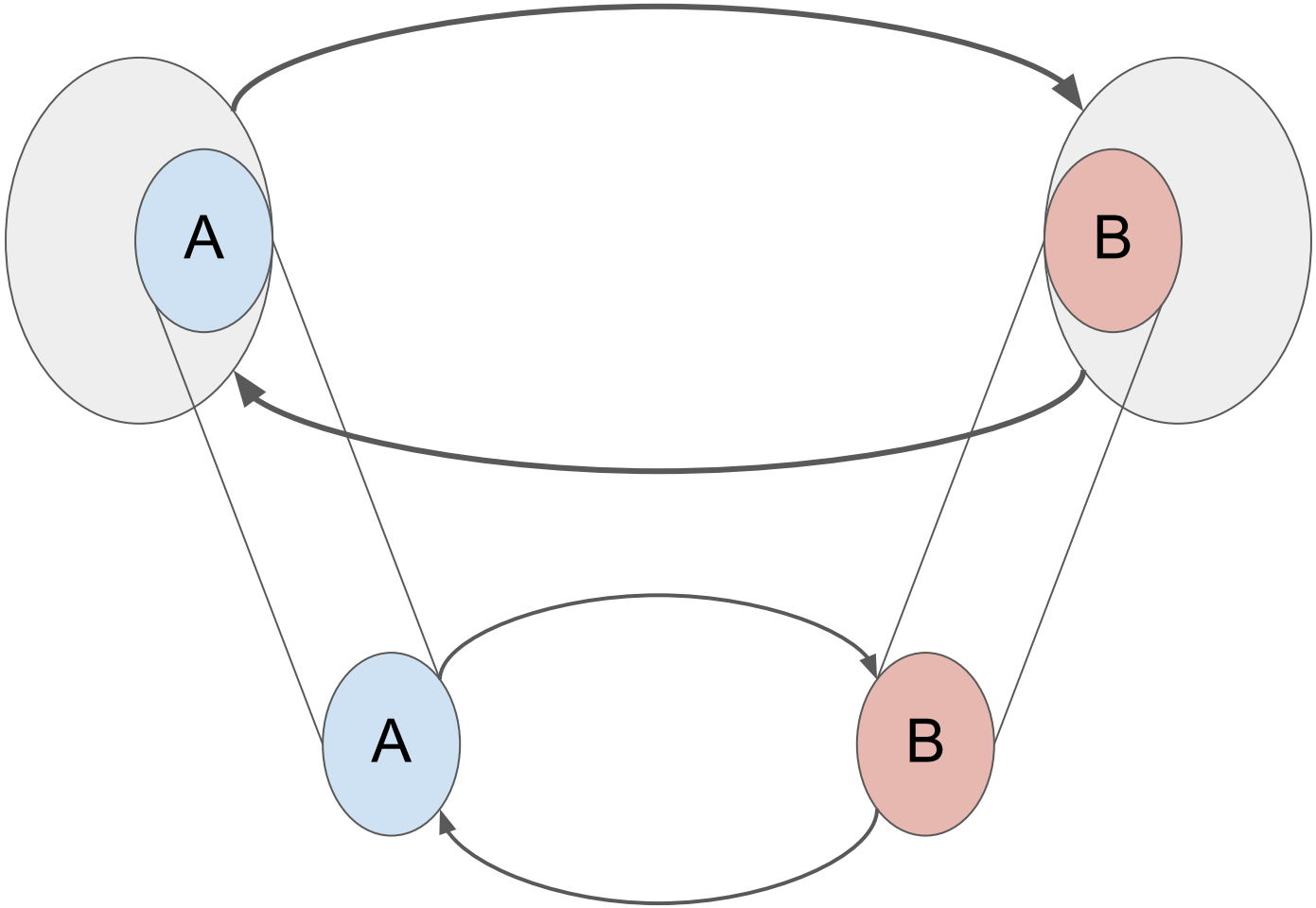
ἡμων ↔ ημων + aspiration

αὐτου ↔ αυτου

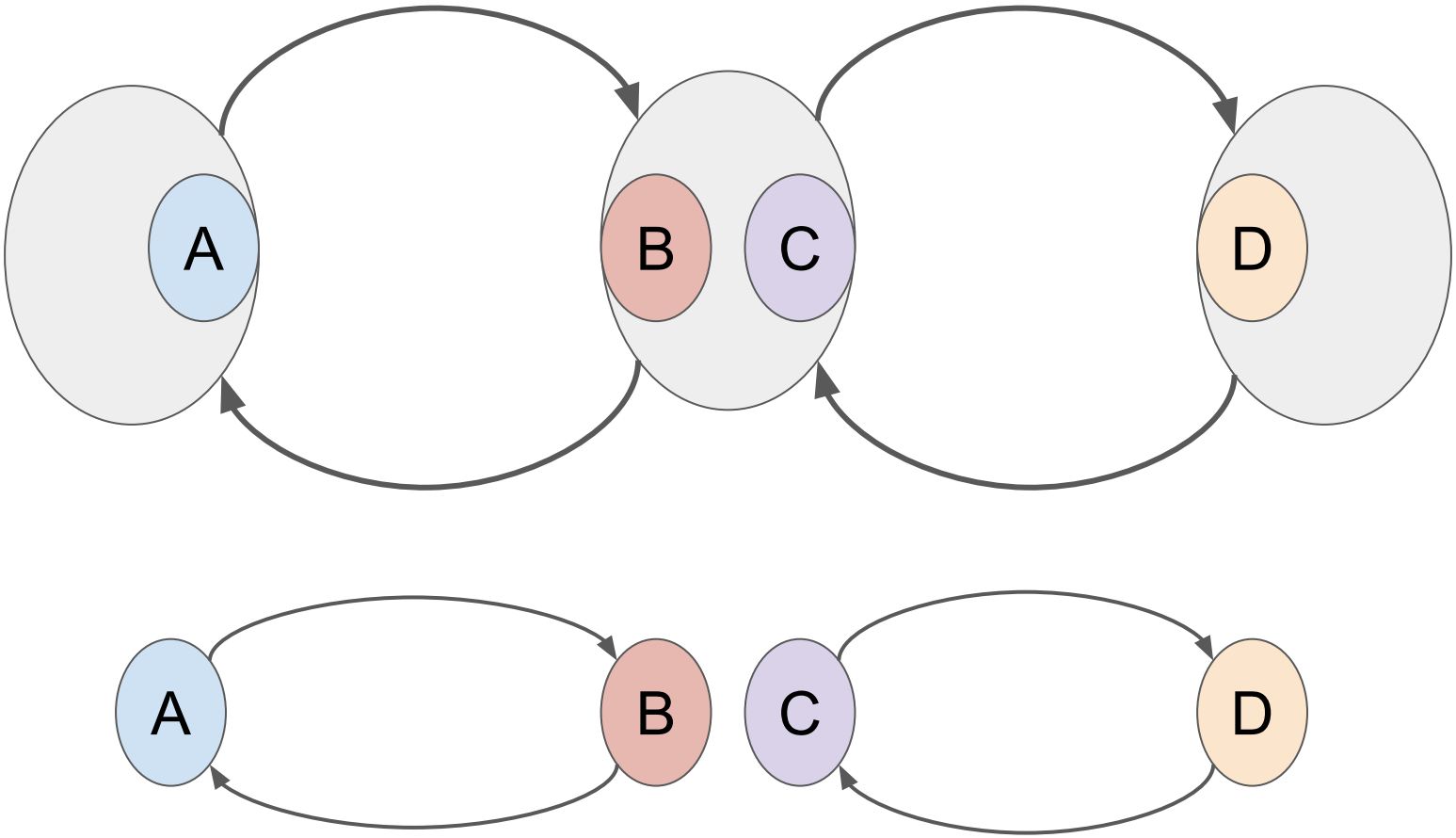
κἀγω ↔ καγω + crasis

Αββα ↔ Αββα + unmarked

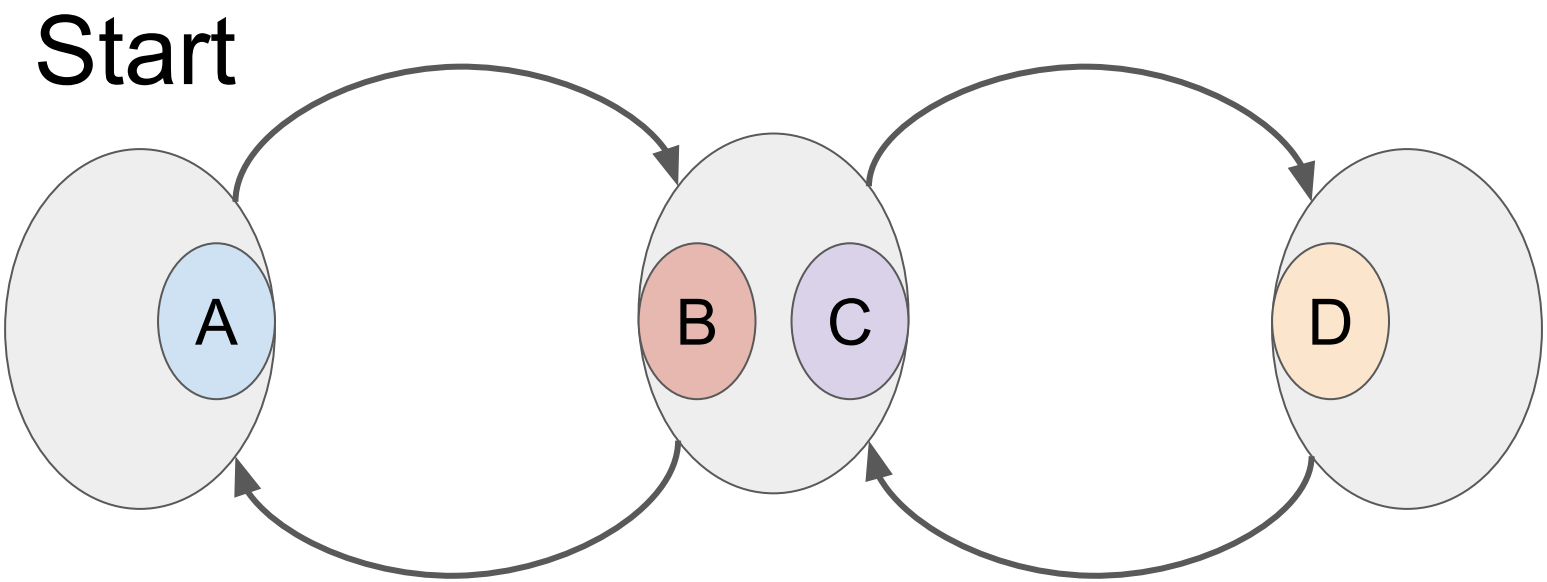
Contextualization of analysis/transformation



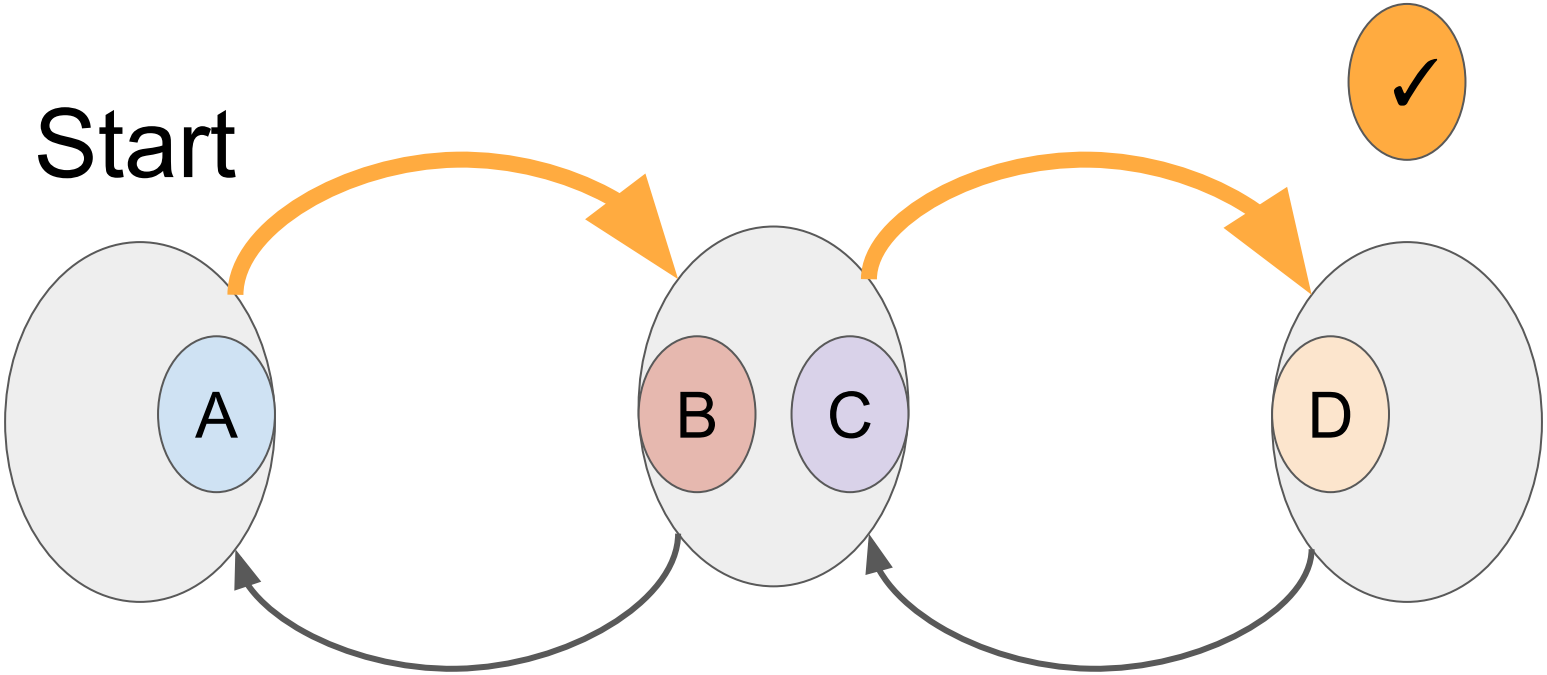
Composition of analyses/transformations



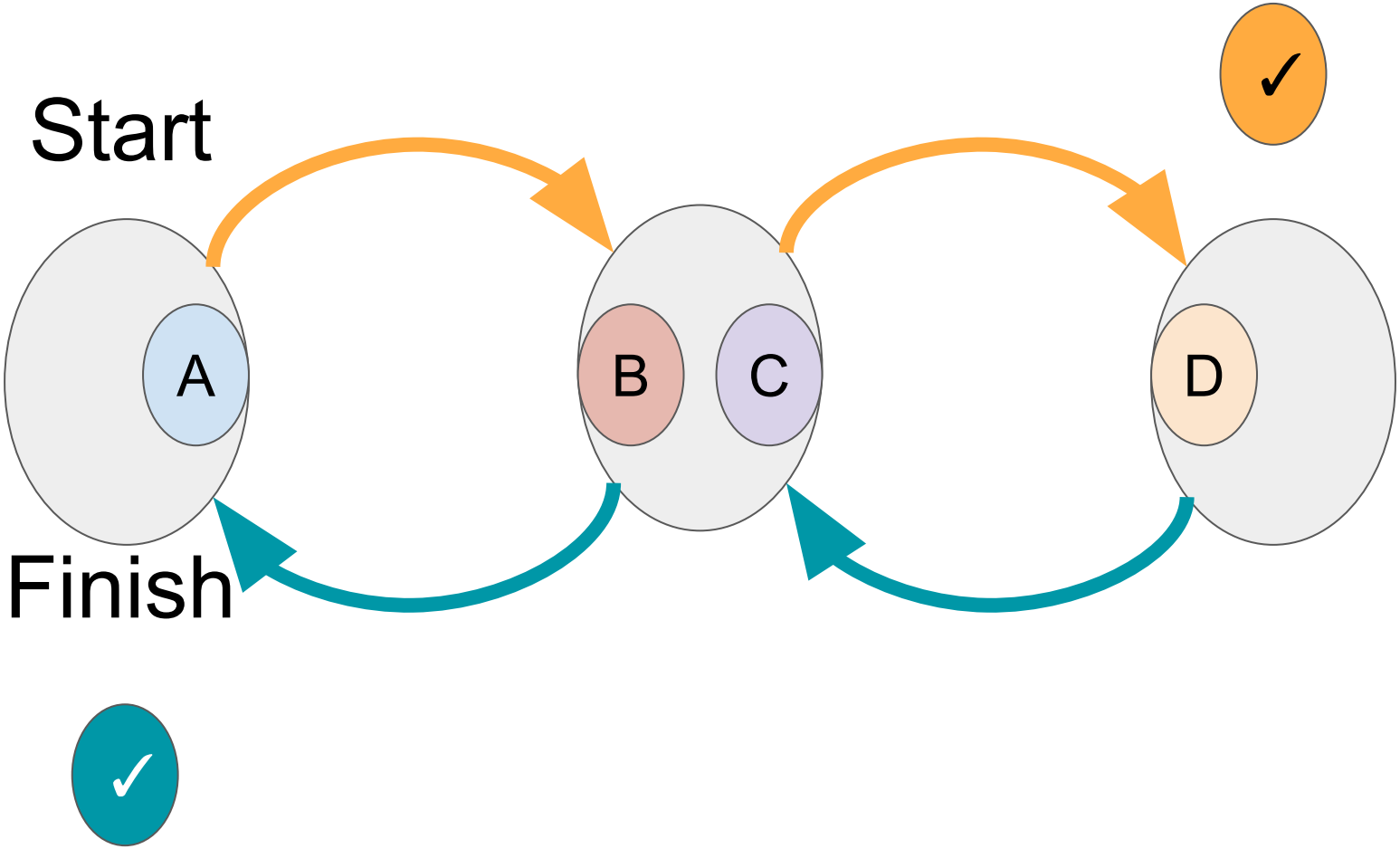
Validation analyses/transformations



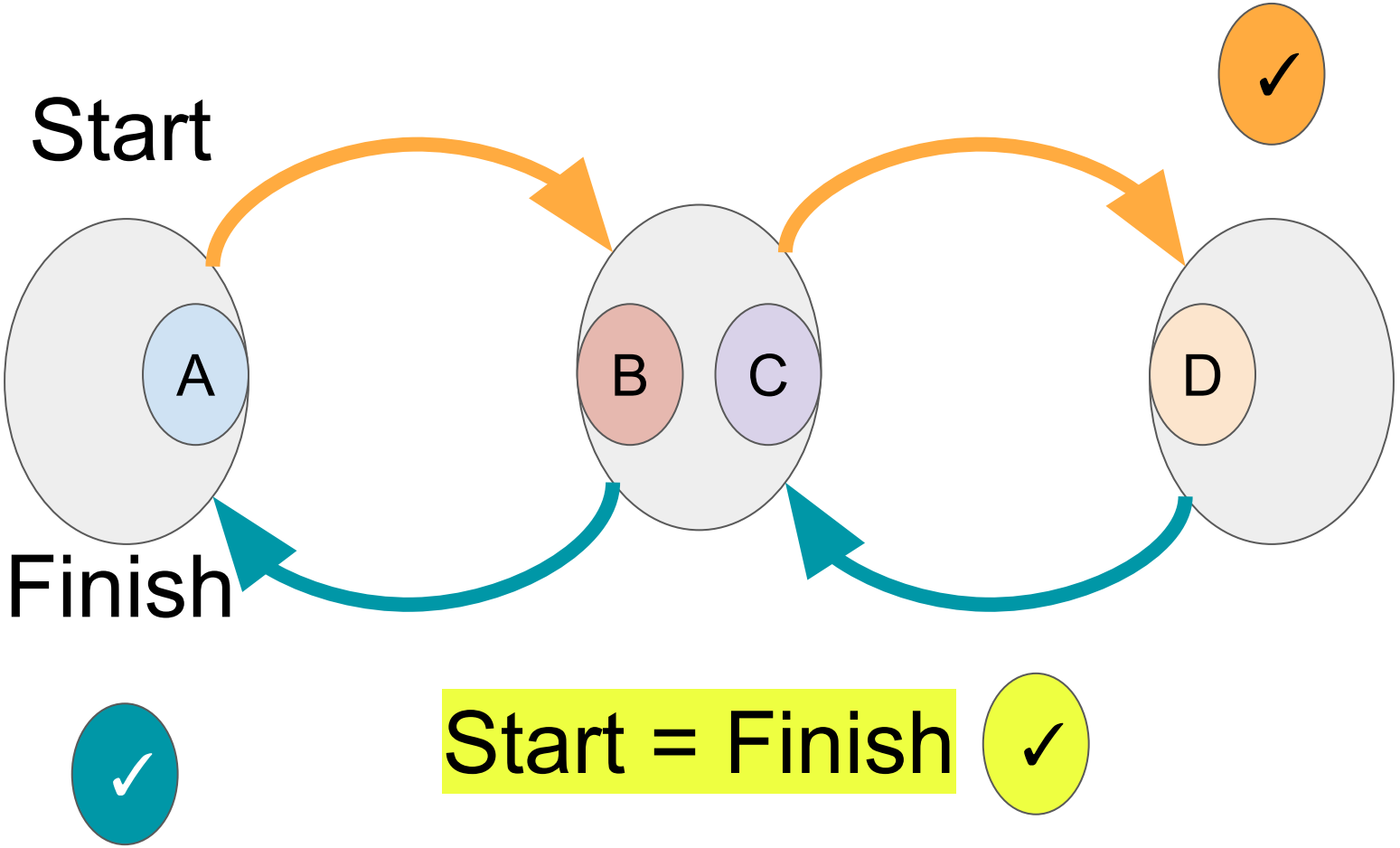
Validation analyses/transformations: 1/3



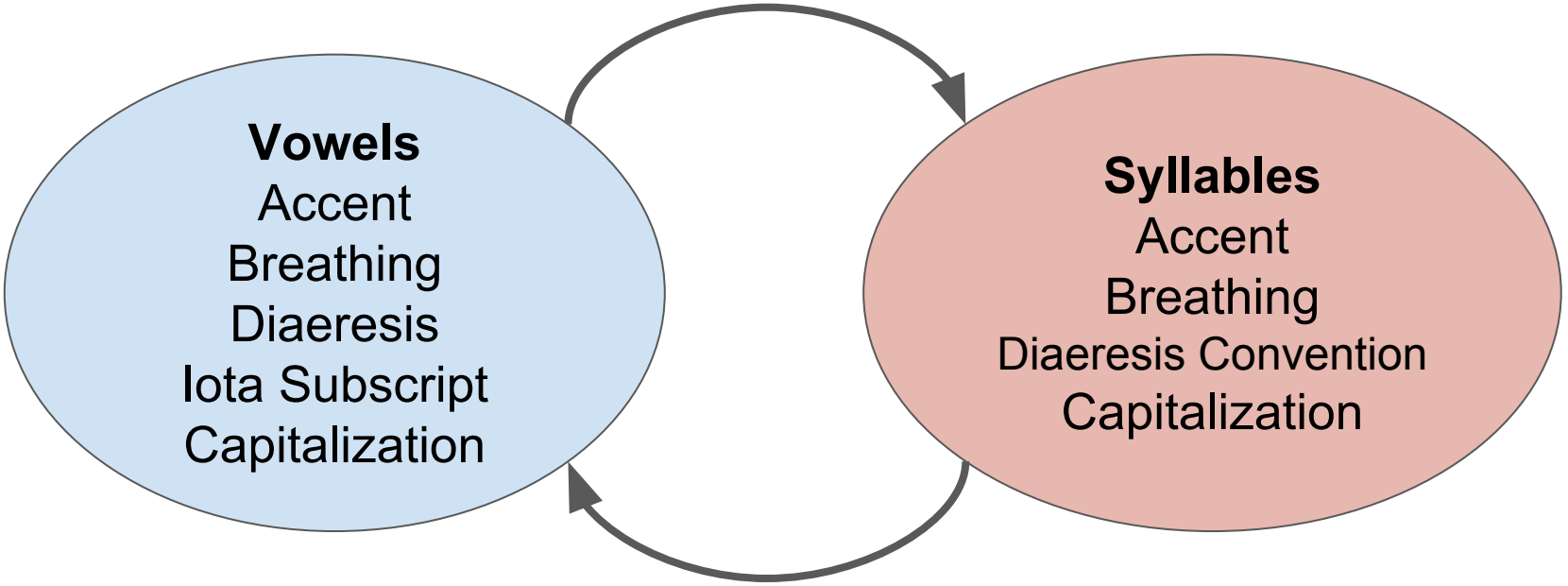
Validation analyses/transformations: 2/3



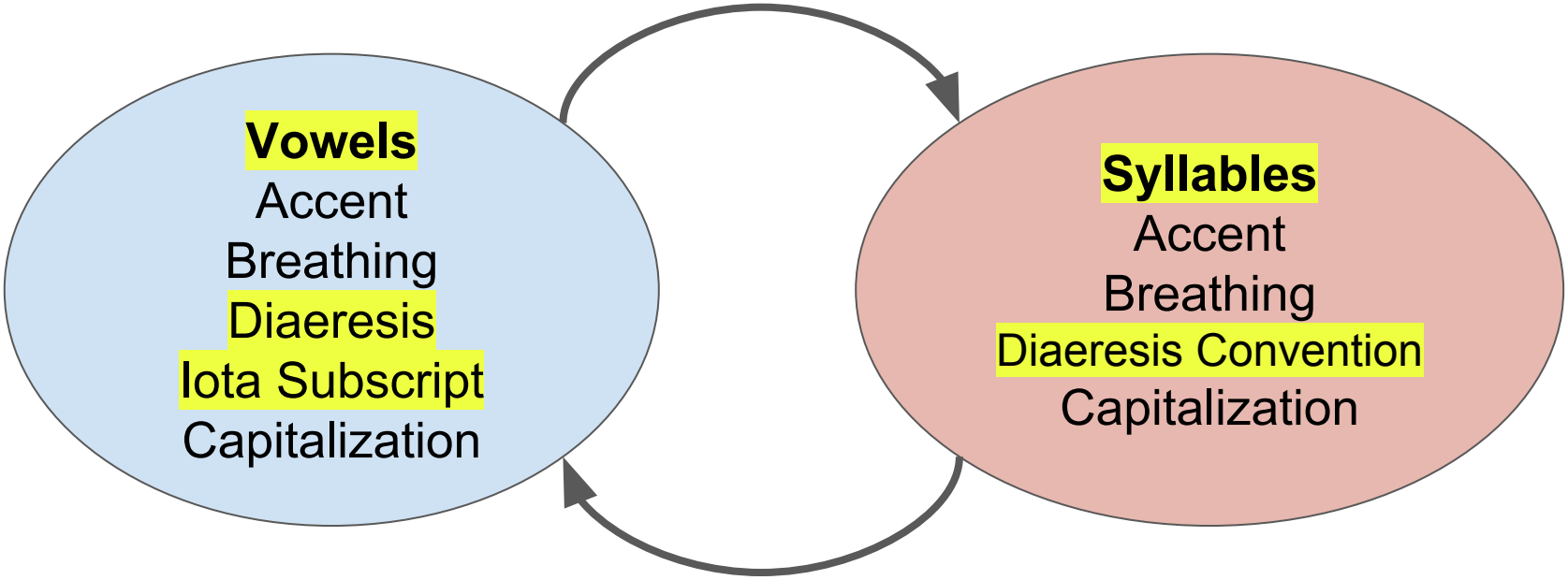
Validation analyses/transformations: 3/3

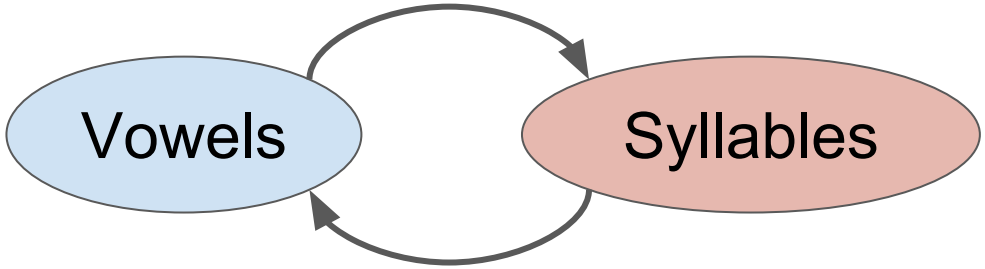


Example: Vowels ↔ Syllables



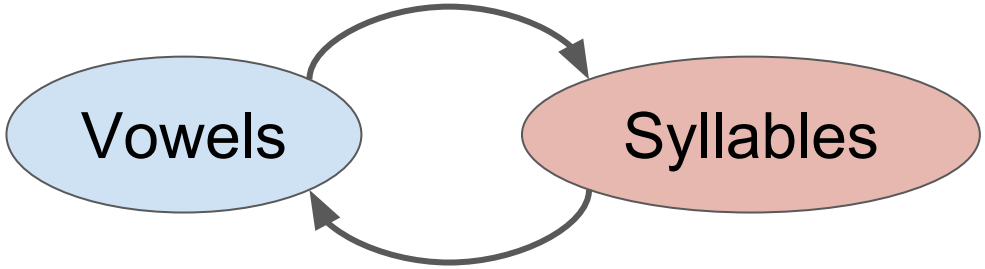
Example: Vowels ↔ Syllables





Rule: Join vowels to form diphthongs

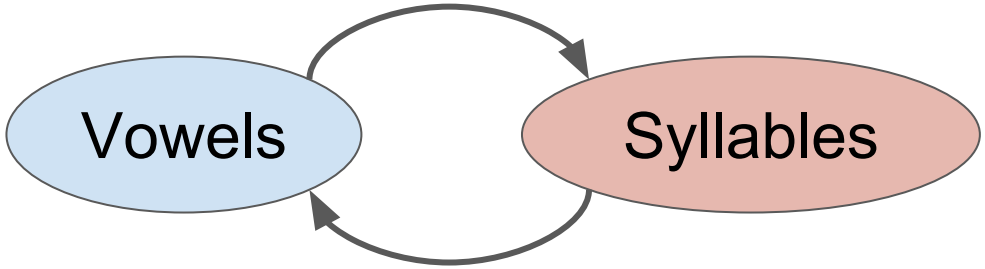
αιου ↔ αι + ου δικ**αίου**ς ↔ δικ**αί + ου**ς



Rule: Join ambiguous diphthongs on the right—no diaeresis

αυι ↔ α + υι (not αυ + ι) Δ**αυί**δ ↔ Δ**α** + **υί**δ (not Δ**αυ** + **ί**δ)

ευι ↔ ε + υι (not ευ + ι) Νιν**ευῖ**ται ↔ Νιν**ε** + **υῖ**ται (not Νιν**ευ** + **ῖ**ται)



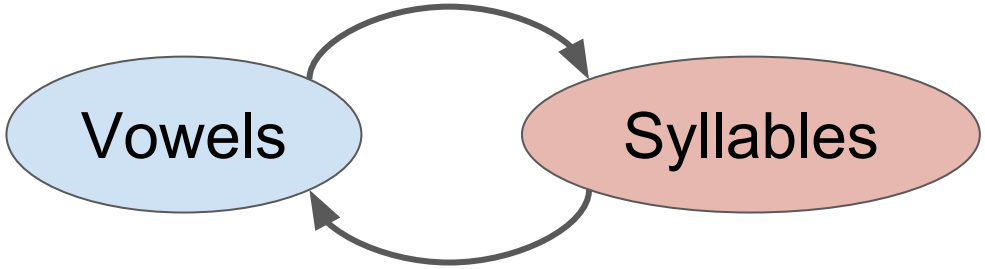
Rule: Vowels with an iota subscript are a syllable

ιῃ ↔ ι + ῃ πίῃ ↔ πί + ῃ

ἰδί + ᾳ ↔ ἰδί + ᾳ

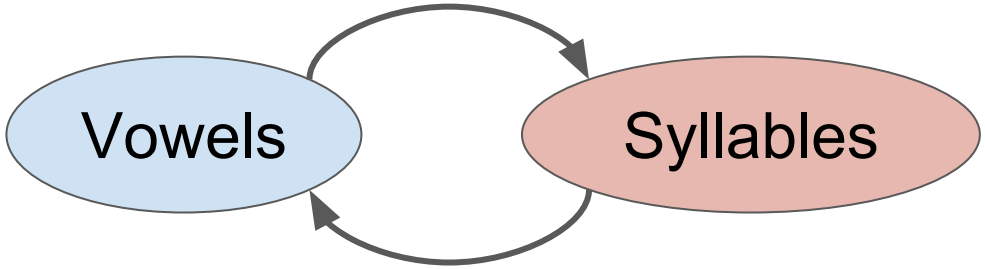
παλα + ι + ῷ ↔ παλαι + ῷ

εἰσῄ + ε + ι ↔ εἰσῄ + ει



Rule: Capital vowels with breathing on first vowel are an iota subscript syllable

Ἅι ↔ ᾷ	**Ἅι**δου ↔ **ᾷ** + δου
Ἠι ↔ ᾐ	**Ἠι**όνα ↔ **ᾐ** + όνα
Ὠι ↔ ᾠ	**Ὠι**δείῳ ↔ **ᾠ** + δείῳ



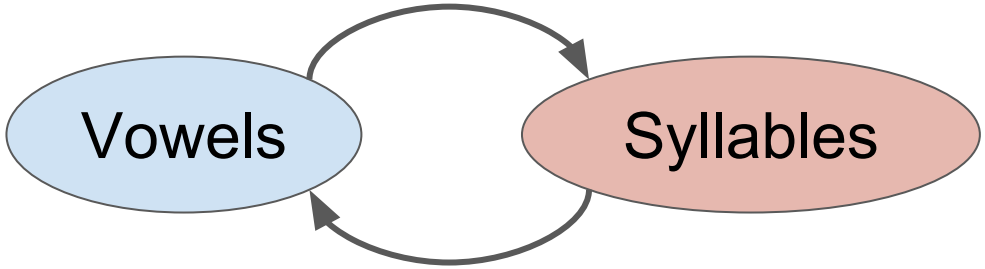
Rule: Diaeresis distinguishes syllables

Diaeresis Convention: Use diaeresis to distinguish syllables

ΰϊ ↔ ΰ + ι ἰσχ**ΰϊ** ↔ ἰσχ**ΰ + ι**

οἰσ**ΰϊ**να ↔ οἰσ**ΰ + ι**να

ώϊ ↔ ώ + ι πρ**ώϊ**μος ↔ πρ**ώ + ι**μος



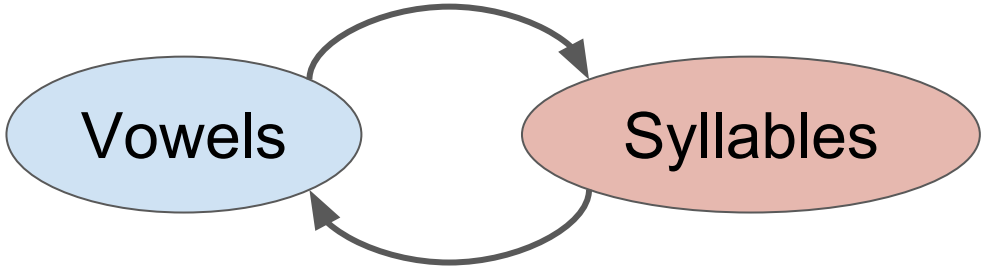
Rule: Accent on first vowel of potential diphthong indicates separate syllables

Diaeresis Convention: Accent can distinguish syllables

ύι ↔ ύ + ι ἰσχ**ύι** ↔ ἰσχ**ύ + ι**

οἰσ**ύι**να ↔ οἰσ**ύ + ι**να

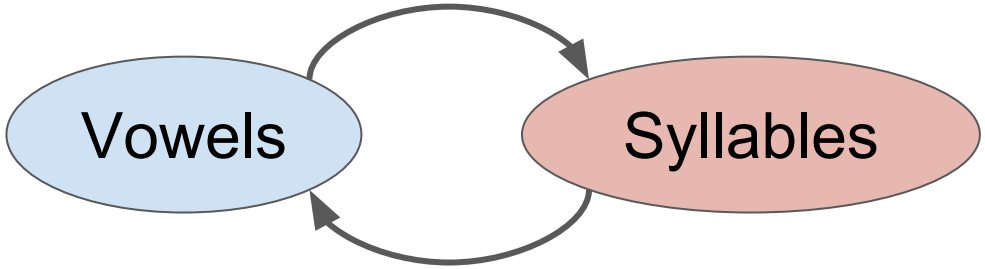
ώι ↔ ώ + ι πρ**ώι**μος ↔ πρ**ώ + ι**μος



Rule: Breathing on first vowel of potential diphthong indicates separate syllables

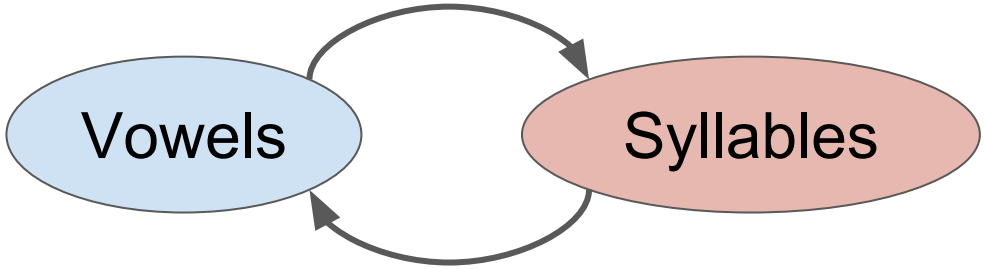
Diaeresis Convention: Breathing and/or accent can distinguish syllables

ἀί ↔ ἀ + ί	**ἀί**διον ↔ **ἀ + ί**διον
ἐυ ↔ ἐ + υ	**ἐυ**ξέστου ↔ **ἐ + υ**ξέστου
ὐι ↔ ὐ + ι	**ὐι**κὸν ↔ **ὐ + ι**κὸν
εὐι ↔ εὐ + ι	**εὐι**ππότην ↔ **εὐι**ππότην



Rule: Diaeresis goes in middle of three syllables

ουϊά ↔ ου + ι + ά Ἀλληλ**ουϊά** ↔ Ἀλληλ**ου + ι + ά**

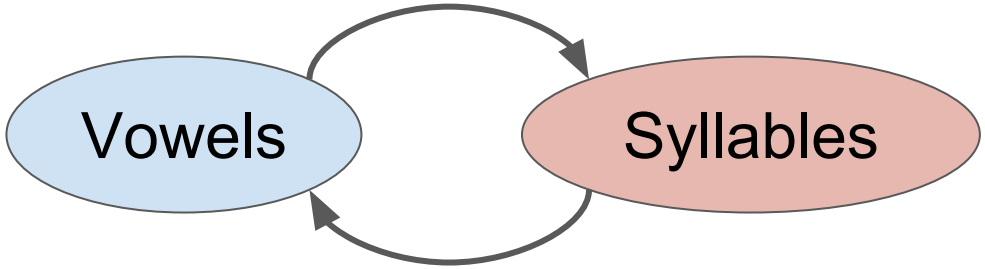


Rule: Vowels that don’t form a diphthong don’t need a diaeresis

Diaeresis Convention: Omit unnecessary diaereses

ηὶ ↔ η + ὶ νη**ὶ** ↔ νη **+ ὶ**

ιὶ ↔ ι + ὶ Δι**ὶ** ↔ Δι **+ ὶ**

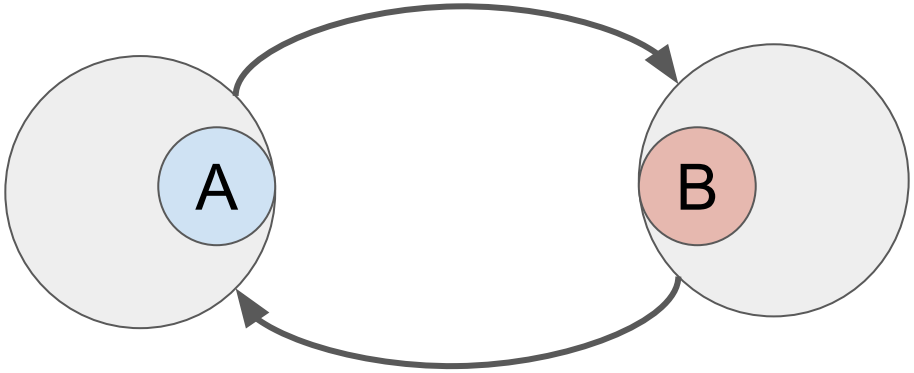


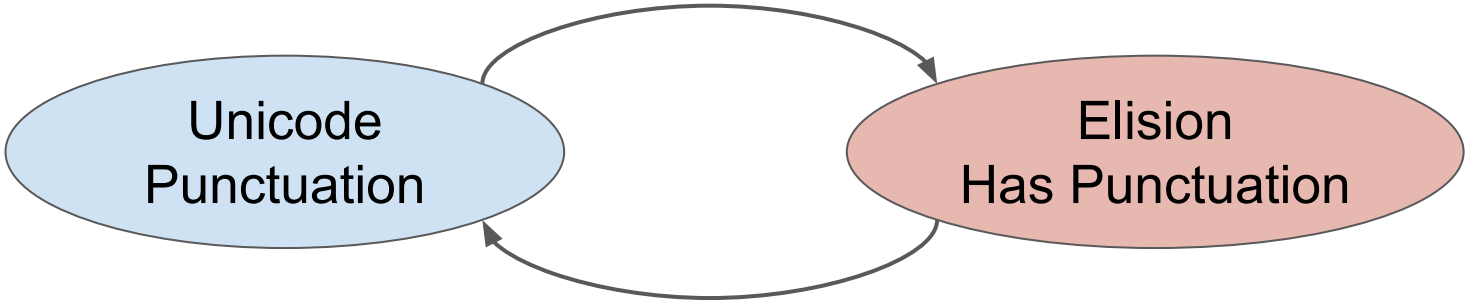
Rule: Use a diaeresis on a second iota or upsilon

Diaeresis Convention: Use diaeresis on all iota/upsilon vowel pairs

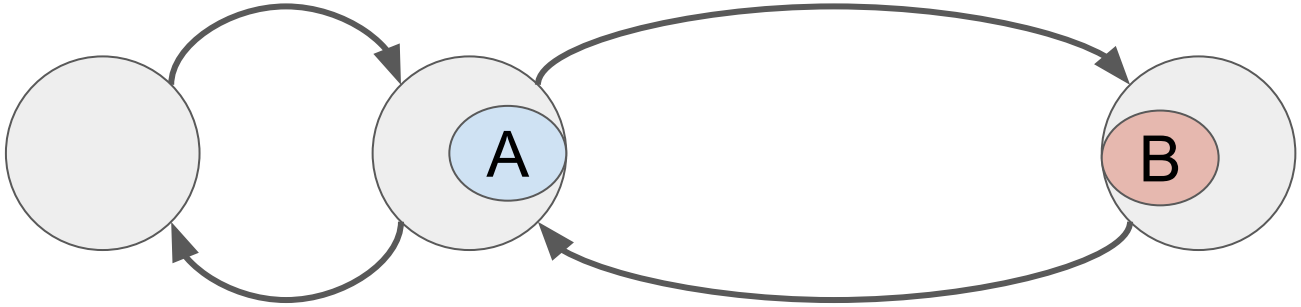
ωϋ ↔ ω + υ	Μ**ωϋ**σῆς ↔ Μ**ω + υ**σῆς
ιϋ ↔ ι + υ	δ**ιϋ**λίζοντες ↔ δ**ι + υ**λίζοντες
ιϊ ↔ ι + ι	δ**ιϊ**σχυρίζετο ↔ δ**ι + ι**σχυρίζετο
οιΐ́α ↔ οι + ί + α	εὐπ**οιΐ́α**ς ↔ εὐπ**οι + ί + α**ς

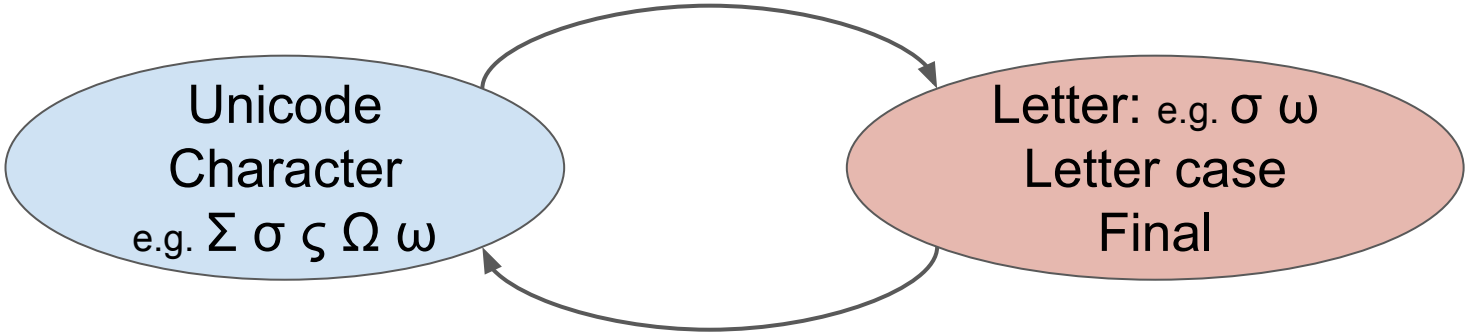
Script Stages: Elision, Punctuation



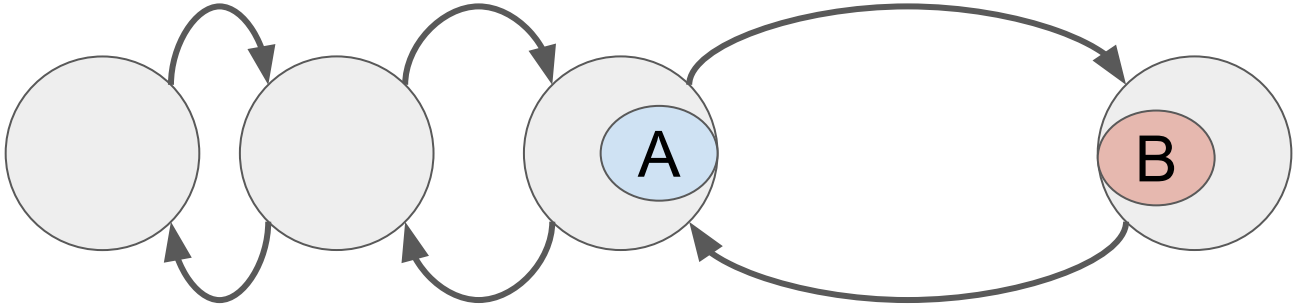


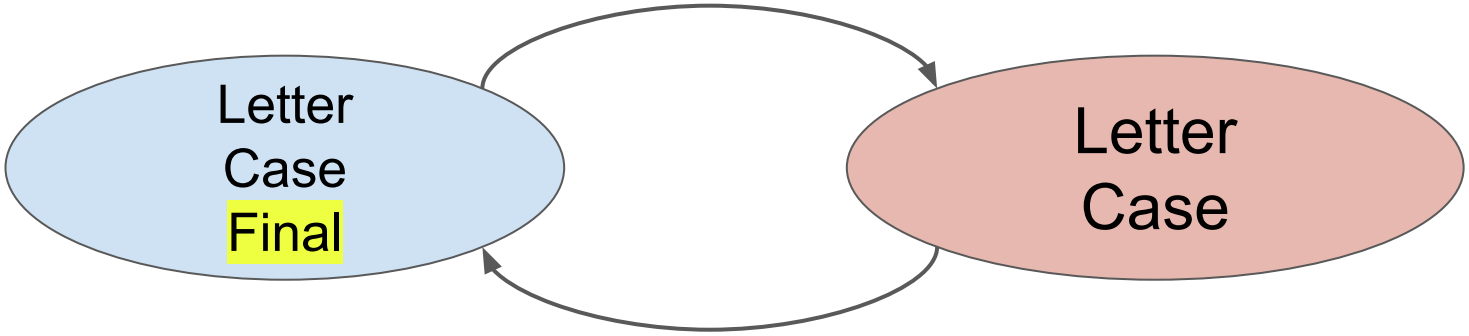
Script Stages: Letter, Case, Final Forms



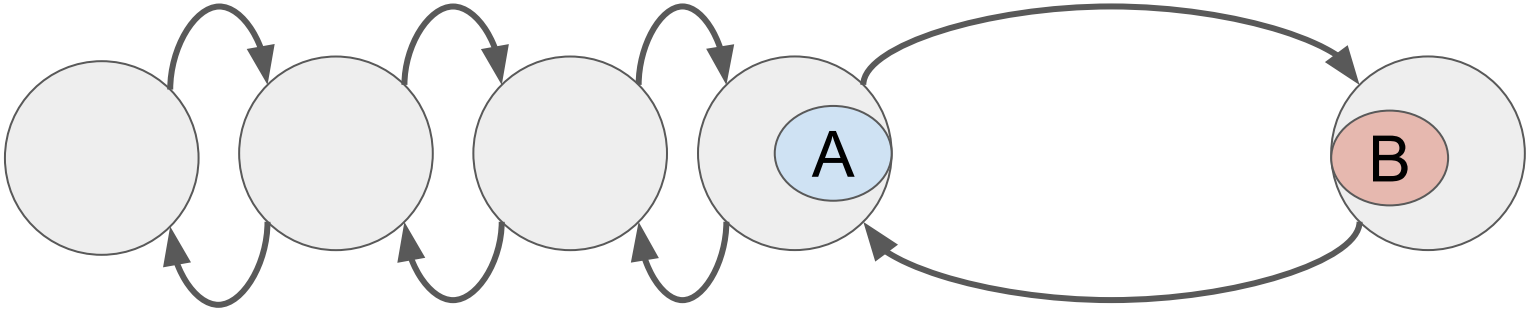


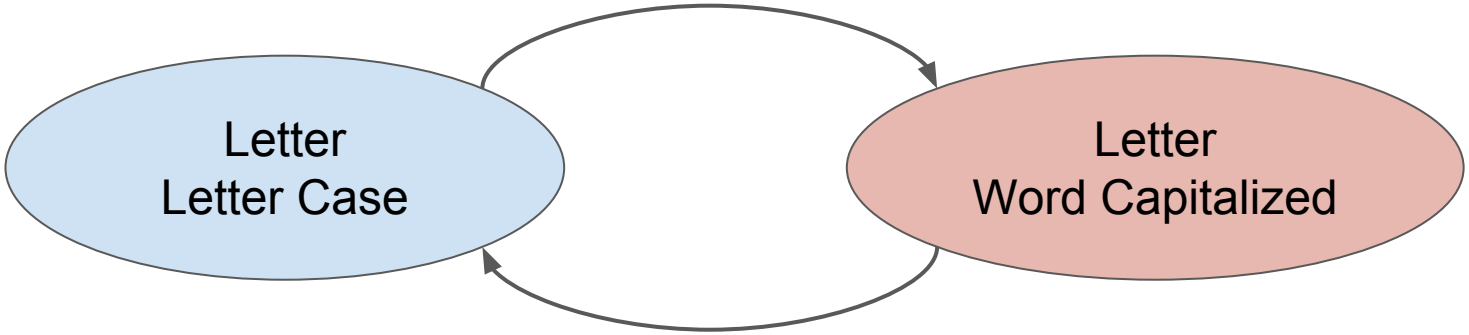
Script Stages: Validate Final Forms



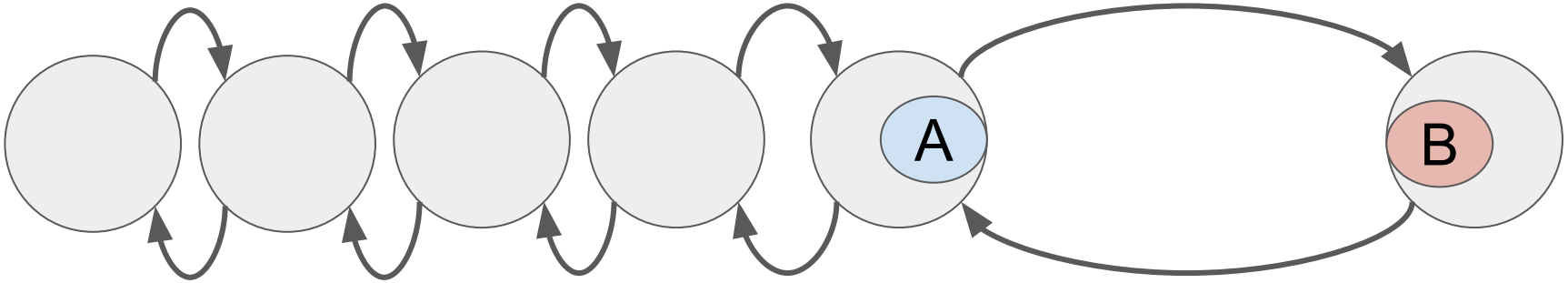


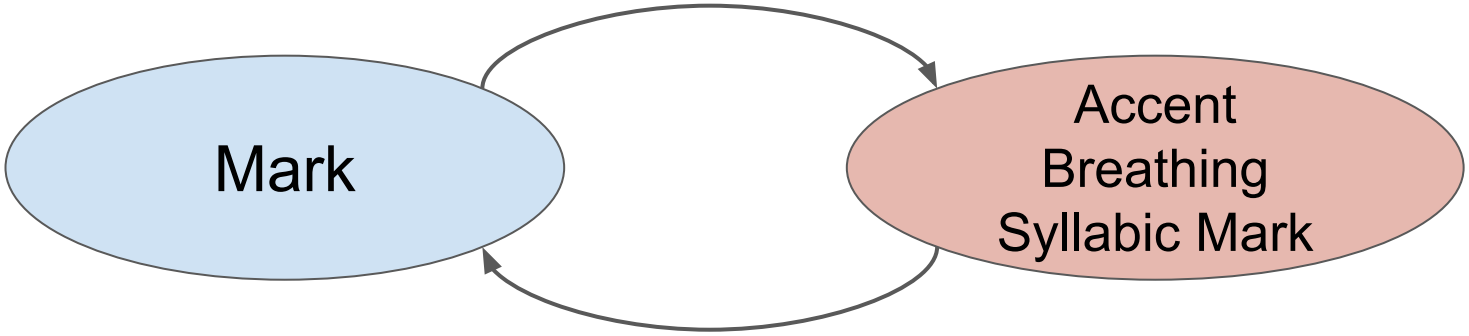
Script Stages: Word Capitalization



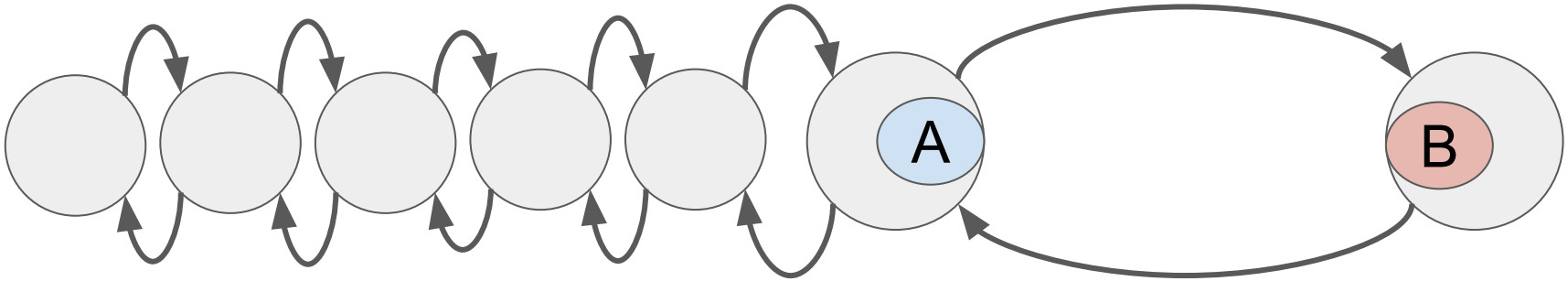


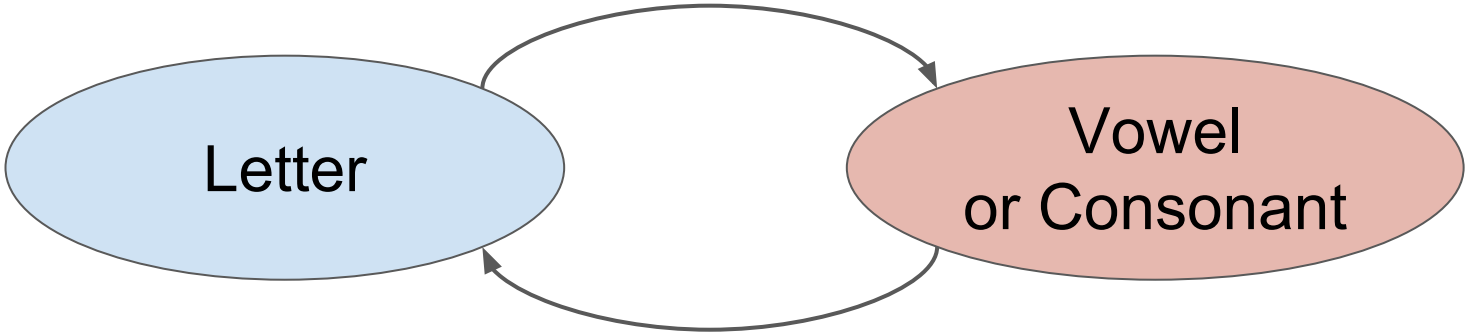
Script Stages: Marks: Accent, Breathing…



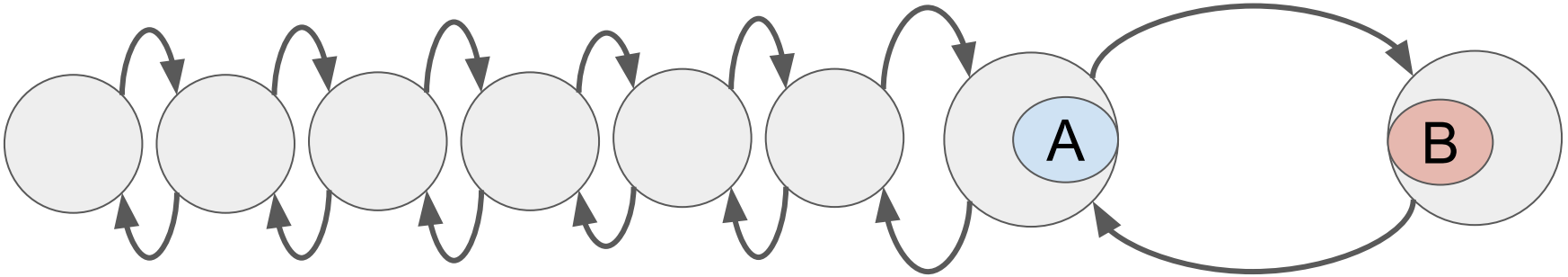


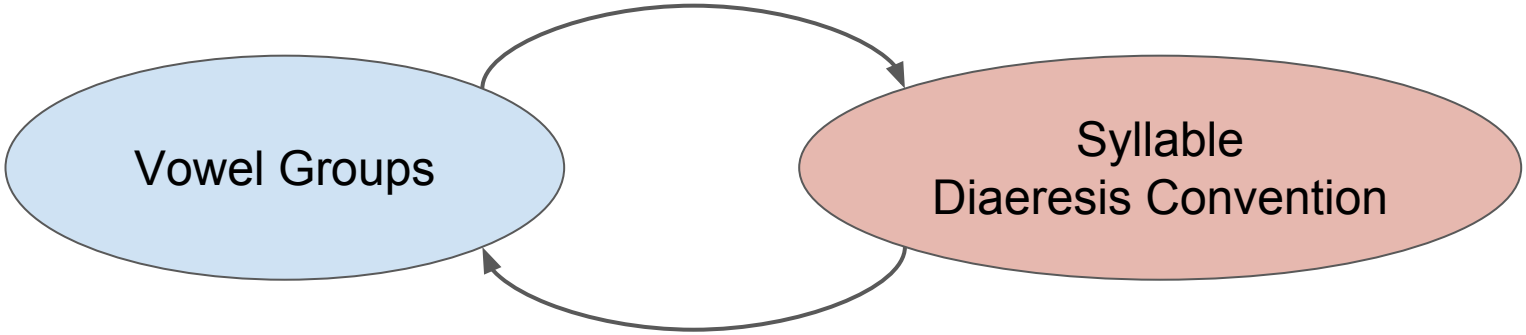
Script Stages: Vowel, Consonant



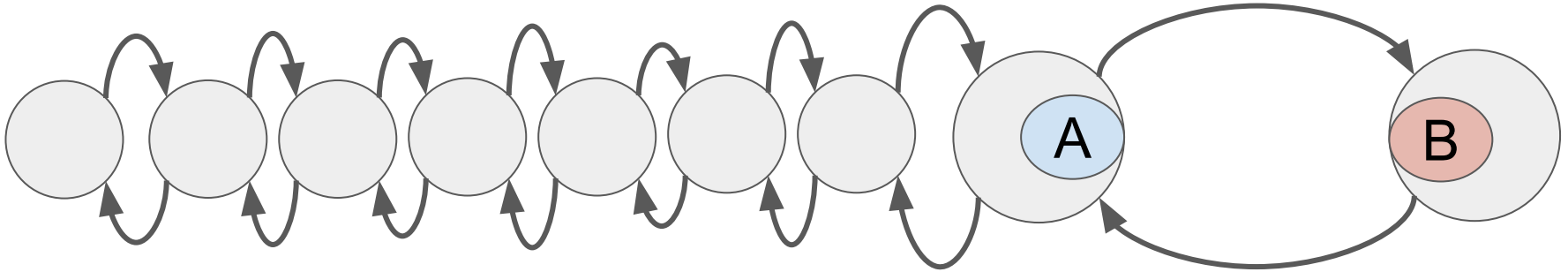


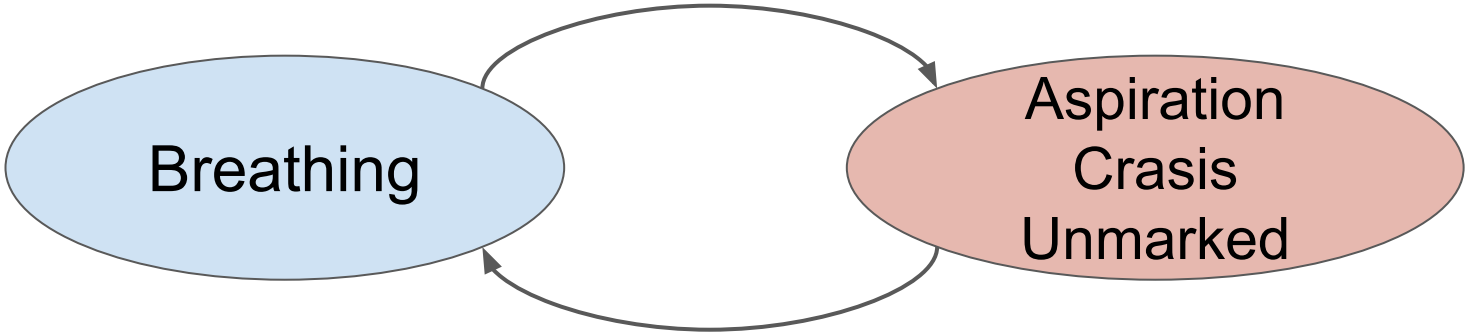
Script Stages: Syllable



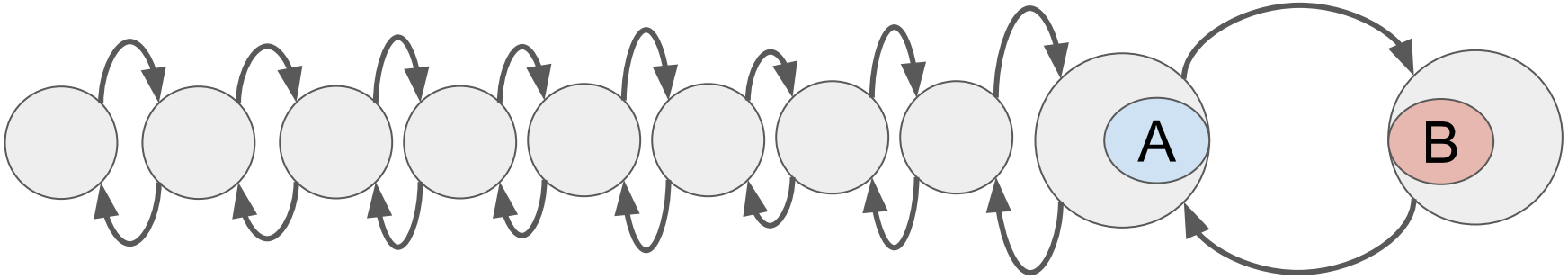


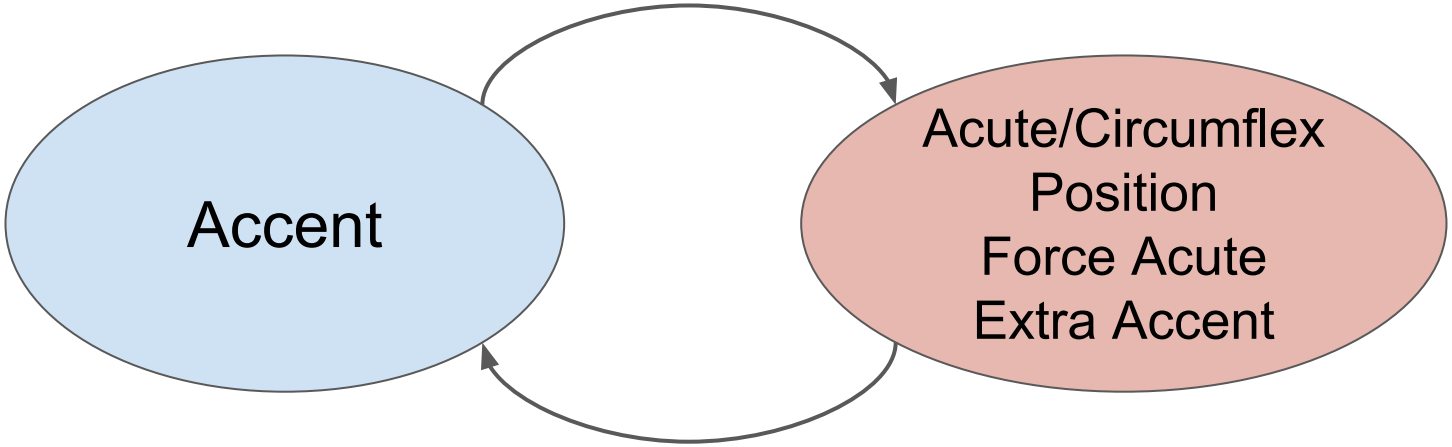
Script Stages: Breathing





Script Stages: Accent

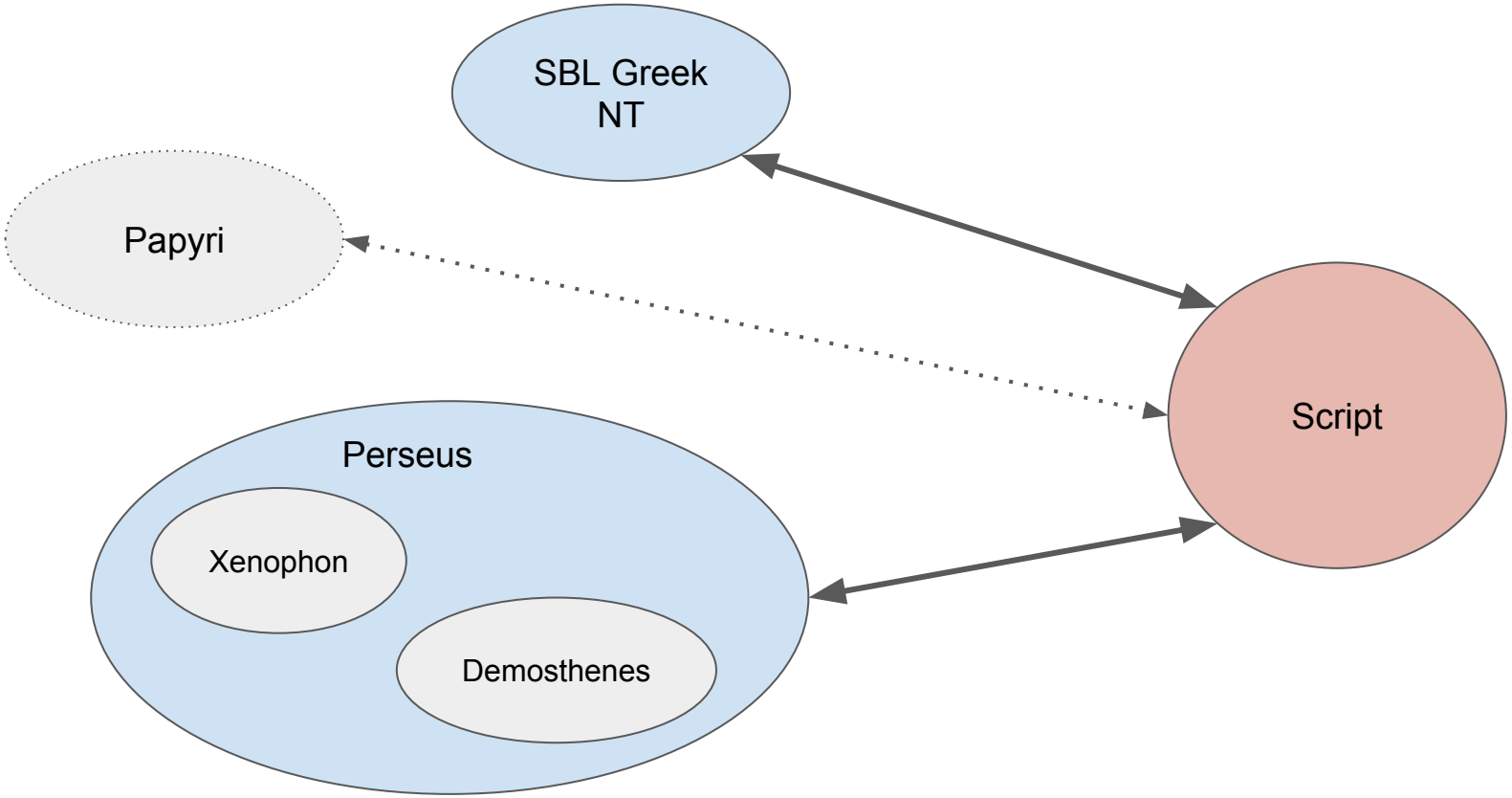




Script: Summary of Word Properties

1. Initial Aspiration: yes / no
2. Syllables, each of which consist of:
   a. Initial consonants
   b. Single vowel, diphthong, or improper diphthong
3. Final Consonants
4. Optional Accent with properties:
   a. Acute or circumflex
   b. Position: Ultima, penult, or antepenult
   c. Force acute: yes / no
   d. Extra accent: yes / no
5. Diaeresis convention
   a. Accent/breathing distinguishes syllables: yes / no
   b. Always use diaeresis with ι,υ: yes / no
6. Crasis: yes with position / no
7. Elision: yes / no
8. Unmarked: yes / no
9. Capitalized: yes / no
10. Has punctuation: yes / no

Script: Primary Sources



Error Correction

SBL Greek NT

- Fixed 12 grave accents with punctuation
  - e.g. αὐτὸν. → αὐτόν.
- Fixed 4 double accent errors
  - e.g. θέλοντές. → θέλοντες.

Error Correction

SBL Greek NT

- Fixed 12 grave accents with punctuation
- Fixed 4 double accent errors

Perseus

- Fixed 1 grave accent error.
  - δὲ, → δέ,
- Final sigmas
  - †δραχμῆσ† Λακεδαιμονίουσς
- Parentheses errors from beta code errors
  - e.g. ἑἶχε δὲ καλόν’ → (εἶχε δὲ καλόν)
- XML errors
  - e.t. ltἂνgt → <ἂν> → <ἂν>
- Paragraph error—no Greek text
- 10+ pull requests, ~50 corrections

Model/Rule Correction

Rule: Use ς for σ at the end of a word (final sigma)

Model/Rule Correction

Rule: Use ς for σ at the end of a word (final sigma)

But… καὶ ὃσ’ εἰς ἡμᾶς ἐβλασφήμησαν

Model/Rule Correction

Rule: Use ς for σ at the end of a word (final sigma)

But… καὶ ὃσ’ εἰς ἡμᾶς ἐβλασφήμησαν

Rule: Circumflex cannot go on the antepenult

Model/Rule Correction

Rule: Use ς for σ at the end of a word (final sigma)

But… καὶ ὅσ’ εἰς ἡμᾶς ἐβλασφήμησαν

Rule: Circumflex cannot go on the antepenult

But… ᾧτινι, οἵάπερ

(cf. Smyth 186)

Model/Rule Correction

Rule: Use ς for σ at the end of a word (final sigma)

But… καὶ ὅσ’ εἰς ἡμᾶς ἐβλασφήμησαν

Rule: Circumflex cannot go on the antepenult

But… ᾧτινι, οἵἁπερ

(cf. Smyth 186)

Rule: Crasis only occurs on first syllable (e.g. κἀγὼ)

Model/Rule Correction

Rule: Use ς for σ at the end of a word (final sigma)

But… καὶ ὅσ’ εἰς ἡμᾶς ἐβλασφήμησαν

Rule: Circumflex cannot go on the antepenult

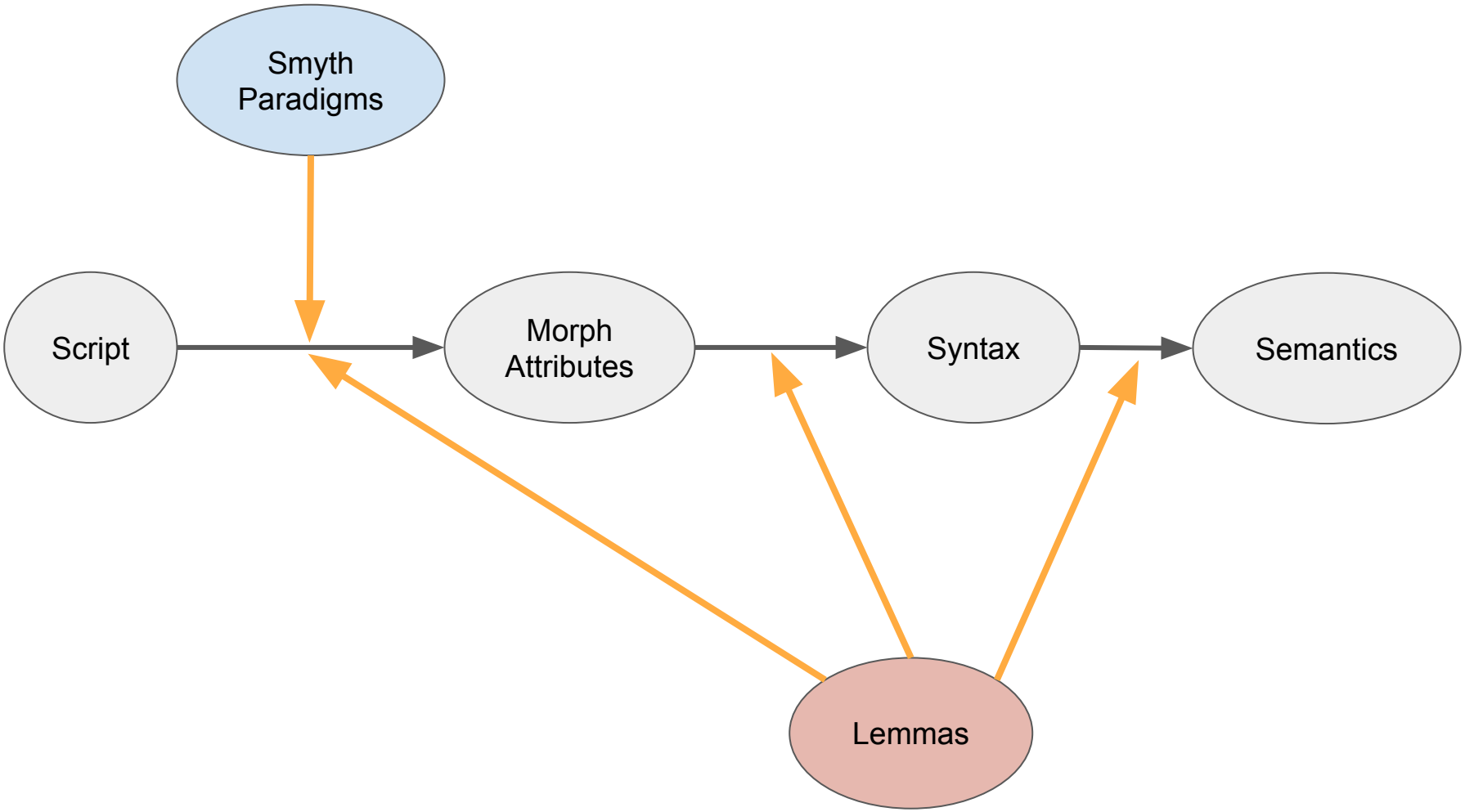
But… ᾧτινι, οἵἅπερ

(cf. Smyth 186)

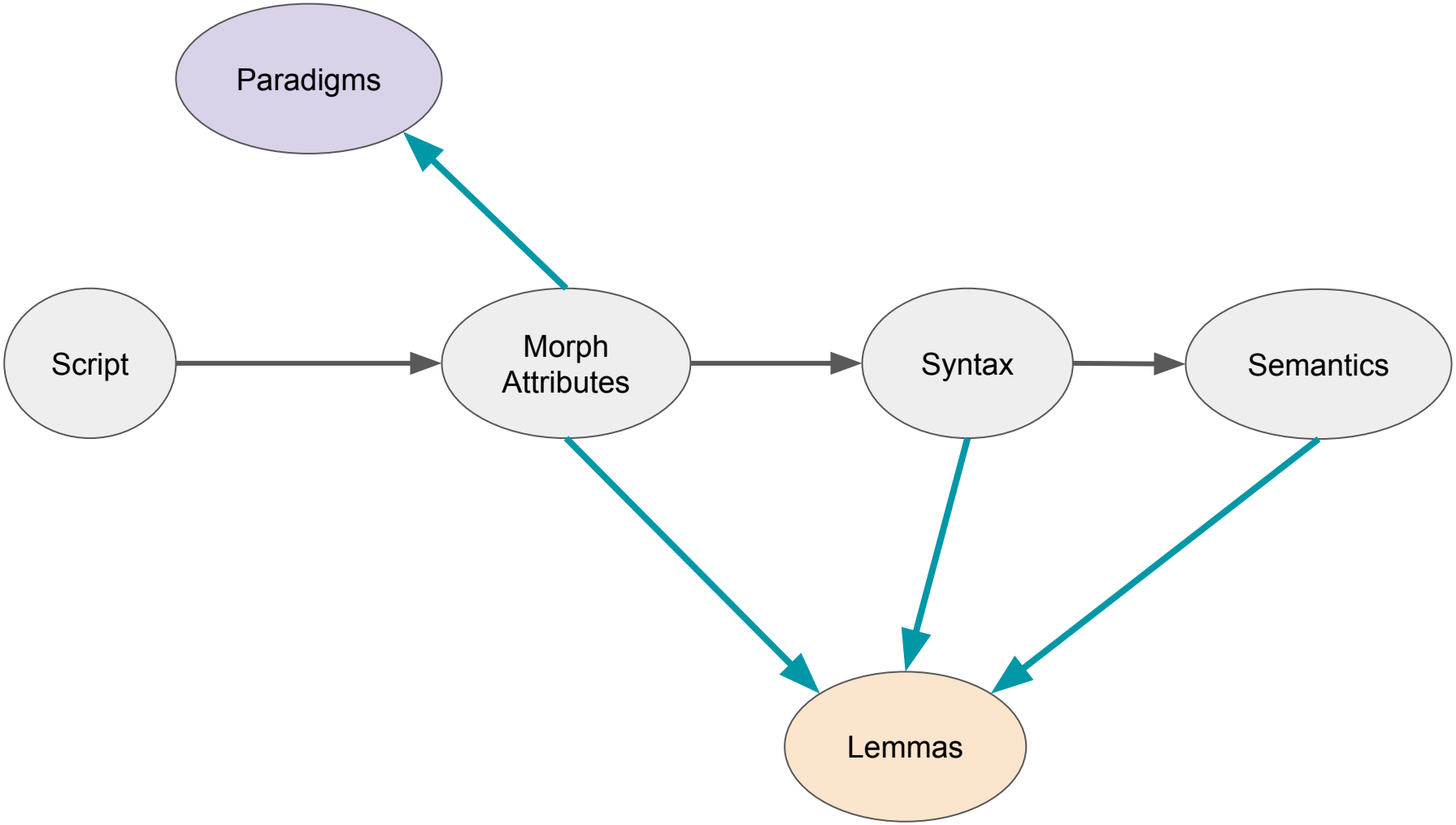
Rule: Crasis only occurs on first syllable (e.g. κἀγὼ)

But… καλοκἀγαθία, ἐγᾦμαι, μεντἂν

Incorporating Existing Analyses



Revising Existing Analyses



Example: First Declension *Feminine* Nouns

216. I. FEMININES

	ἠχώρᾱ (χωπᾱ-)	ἡ νῑκη (νῑκᾱ-)	ἡ φυγή (φυγᾱ-)	ἡ μοῖρα (μοιρᾱ)	ἡ γλῶττα (γλωττᾱ-)	ἡ θάλαττα (θαλαττᾱ)
	land	*victory*	*flight*	*fate*	*tongue*	*sea*
SINGULAR						
Nom.	**χώρᾱ**	**νῑ́κη**	**φυγή**	**μοῖρα**	**γλῶττα**	**φάλαττα**
Gen.	**χώρᾱς**	**νῑ́κης**	**φυγῆς**	**μοίρᾱς**	**γλωιττης**	**φαλάττης**
Dat.	**χώρᾳ**	**νῑ́κῃ**	**φυγῇ**	**μοίρᾳ**	**γλώττῃ**	**φαλάττῃ**
Acc.	**χώρᾱ-ν**	**νῑ́κη-ν**	**φυγή-ν**	**μοῖρα-ν**	**γλῶττα-ν**	**θάλαττα-ν**
Voc.	**χώρᾱ**	**νῑ́κη**	**φυγή**	**μοῖρα**	**γλῶττα**	**θάλαττα**
DUAL						
N. A. V.	**χώρᾱ**	**νῑ́κᾱ**	**φυγᾱ́**	**μοίρᾱ**	**γλώττᾱ**	**φαλάττᾱ**
G. D.	**χώραιν**	**νῑ́καιν**	**φυγαῖν**	**μοίραιν**	**γλώτταιν**	**φαλάτταιν**
PLURAL						
N.V.	**χῶραι**	**νῑ́και**	**φυγαί**	**υοῖραι**	**γλῶτται**	**θάλατται**
Gen.	**χωρῶν**	**νῑκῶν**	**φυγῶν**	**μοιρῶν**	**γλωττῶν**	**φαλαττῶν**
Dat.	**χώραις**	**νῑ́καις**	**φυγαῖς**	**μοίραις**	**γλώτταις**	**θαλάτταις**
Acc.	**χώρᾱς**	**νῑ́κᾱς**	**φυγᾱ́ς**	**μοίρᾱς**	**γλώττᾱς**	**θαλάττᾱς**

Example: *First Declension* Nouns

SUBSTANTIVES

FIRST DECLENSION (STEMS IN **ᾱ**)

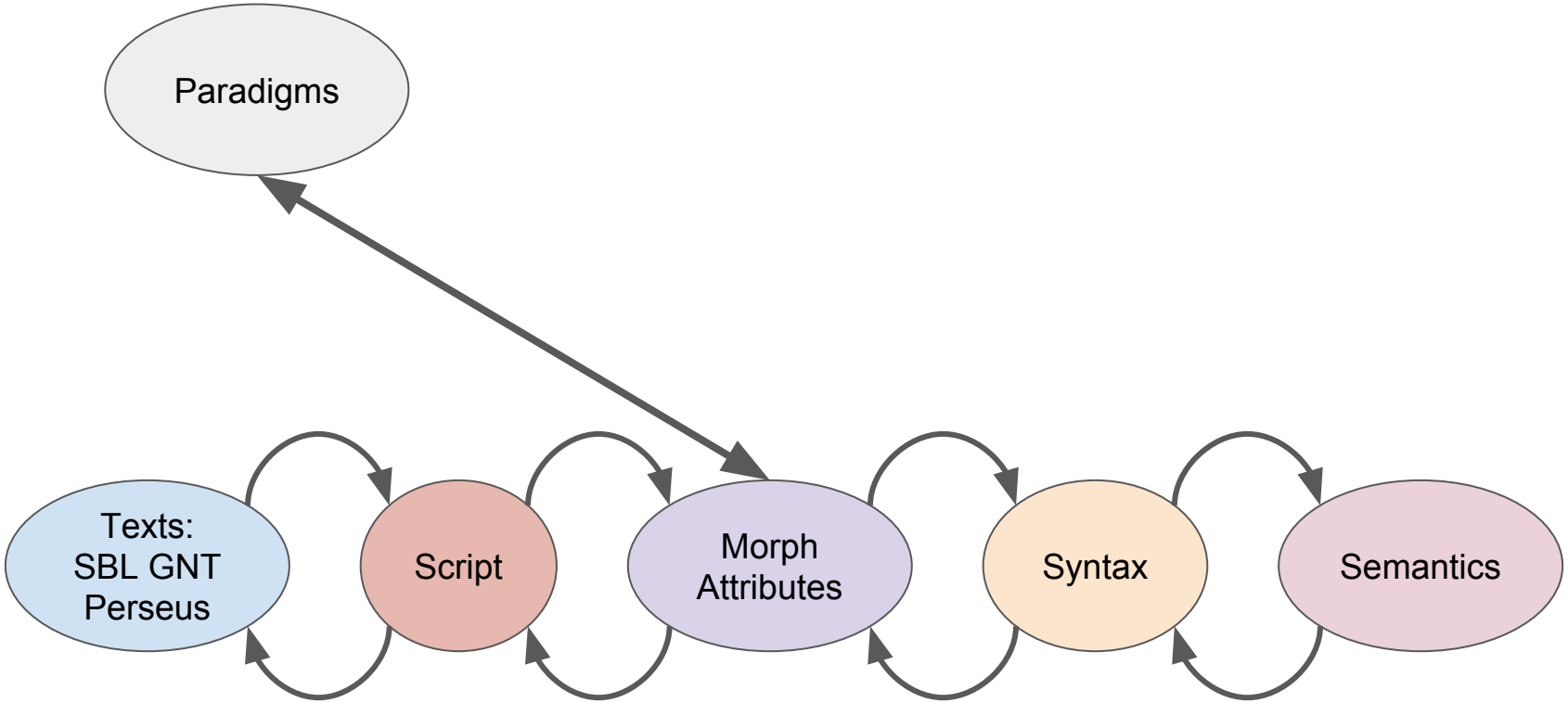
	FEM. SING.		MASC. SING.		MASC. FEM. PL.		MASC.FEM.DUAL	
Nom.	**ᾱ** or **ᾰ**	**η**	**ᾱ-ς**	**η-ς**	**α-ι**	N. A. V.	**ᾱ**	
Gen.	**ᾱ-ς** or **η-ς**	**η-ς**	**ᾱ-ιο** (Hom. **ᾱ-ο**)		**ῶν** (for **έ-ων, ᾱ́-ων**)	G. D.	**α-ιν**	
Dat.	**ᾱ-ι** or **η-ι**	**η-ι**	**ᾱ-ι**	**η-ι**	**α-ις** or **α-ισι** (**ν**)			
Acc.	**ᾱ-ν** or **ᾰ-ν**	**η-ν**	**ᾱ-ν**	**η-ν**	**ᾱς** (for **α-νς**			
Voc.	**ᾱ** or **ᾰ**	**η**	**ᾱ**	**ᾰ** or **η**	**α-ι**			

Example: Nouns

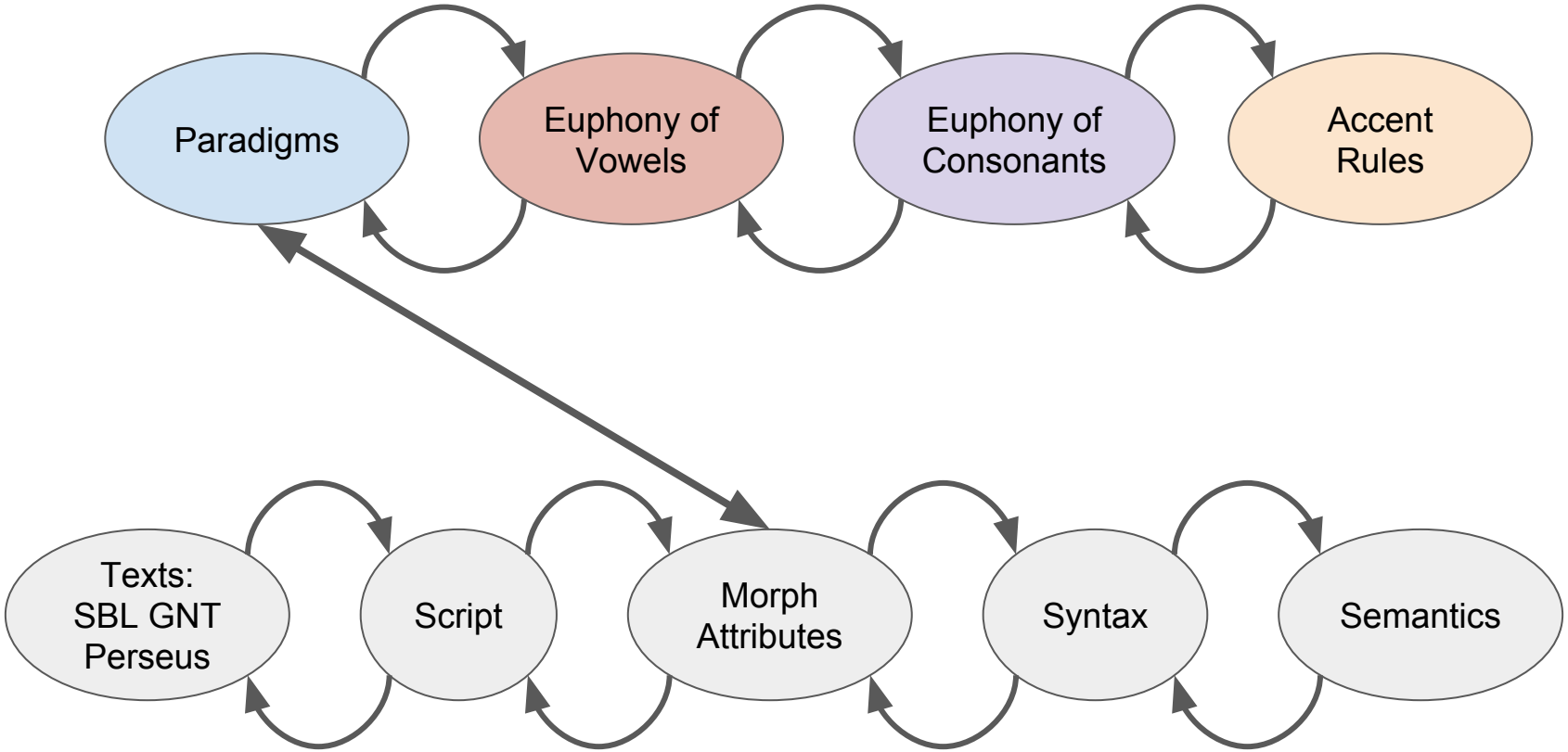
210. CASE ENDINGS OF NOUNS

	VOWEL, DECLENSION		CONSONANT DECLENSION	
	SINGULAR			
	Masc. and Fem.	Neuter	Masc. and Fem.	Neuter
Nom.	-ς or none	-ν	-ς or none	none
Gen.	-ς or -ιο		-ος	
Dat.	-ι		-ι	
Acc.	-ν		-ν or -ᾰ	none
Voc.	none	-ν	none or like Nom.	none
	DUAL			
N. A. V.	none		-ε	
G. D.	-ιν		-οιν	
	PLURAL			
N. V.	-ι	-ᾰ	ε͟ῳ	-ᾰ
Gen.	-ων		-ων	
Dat.	-ις (-ισι)		-σι, -σσι, -εσσι	
Acc.	-νς (-ᾱς)	-ᾰ	-νς, -ᾰς	-ᾰ

Paradigm Analysis



Paradigm Analysis



Example: Word Formation

PART III

FORMATION OF WORDS

839. AGENCY

a. The primary suffixes τᾱ, τηρ, τορ, τρο, ευ, denoting the *agent* or *doer of an action*, are masculine.

1. **τᾱ** (nom. -τή-ς): κρι-τή-ς *judge* (κρῑ́νω *decide*, κρι-), κλέπ-τη-ς *thief* (κλέπ-τ-ω *steal*), ποιη-τή-ς *poet*, i.e. *maker* (ποιέ-ω *make*), αὐλη-τή-ς *flute-player* (αὐλέ-ω *play the flute*), μαθ-η-τή-ς *pupil* (μανθάνω *learn*, μαθ-ε-), ἱκ-έ-τη-ς *suppliant* (ἱκ-νέ-ομαι *come*, ἱκ-).
2. **τηρ** (nom. -τήρ): δο-τήρ *giver* (δί-δω-μι *give*, δο-, δω-), σω-τήρ *saviour* (σῴ-ζω *save*).
3. **τορ** (nom. -τωρ): ῥή-τωρ *orator* (ἐρέω *shall say*, ἐρ-, ῥε-), εἴ-ρη-κα *have spoken*, κτίσ-τωρ *founder* (κτίζω *found*, κτιδ-), σημάντωρ *commander*, poet. (σημαίνω *give a signal*, σημαν-).
4. **τρο** (nom. -τρό-ς): ἰᾱ-τρό-ς *physician* (ἰα-ομαι *heal*).
5. **ευ** (nom, -εύ-ς): γραφ-εύ-ς *writer* (γράφ-ω *write*), τοκ-εύ-ς *father* (τίκτω *beget*, τεκ-).

b. The primary suffixes τριδ, τριᾱ, τειρᾱ, τιδ are feminine.

1. **τριδ** (nom. -τρίς): αὐλη-τρίς *female flute-player*.
2. **τριᾱ** (nom. -τρια): ποίη-τρια *poetess* (late), ψάλ-τρια *female harper* (ψάλλω *play the harp*, ψαλ-).
3. **τειρᾱ** (nom. -τειρα from τερ-ια): σώ-τειρα fem. of σω-τήρ, δό-τειρα fem. of δο-τήρ.
4. **τιδ** (nom. -τις): ἱκ-έ-τις *female suppliant* fem. of ἱκ-έ-της.

Morph Stages

- Decontextualization
  - Elision — How vs When
  - Enclitics as separate words
- Word analysis
  - Enclitic endings
  - Prepositional prefixes
  - Paradigm suffix analysis
  - Augment, reduplication prefix analysis
- Syntactic constituents
  - Preposition + object
  - Conjunctions

Future Work

- More Texts
  - Rest of Perseus
  - Papyri
- More Languages
  - Hebrew
  - Latin
- Deeper analysis
  - Word formation
  - Meter analysis
  - Syntactic constituents
- User Interface
  - Custom language for rules and model
  - Web app