

A background network diagram consisting of numerous gray circles of varying sizes connected by thin gray lines, creating a complex web-like structure.

# Part 2: Family-specific innovations

**Erica Biagetti**

University of Pavia

SunoikisisDC Summer 2022, Session 1  
WordNets for Ancient Indo-European Languages

# In this section

- **Innovations** we have introduced to the WordNet architecture, in order to:
  1. Better capture the **polysemy of words**, including their figurative metaphorical and metonymic uses
    - ☞ Cognitive Linguistics
  2. Account for Indo-European **family-specific features**
    - ☞ Diachronic and stylistic metadata
    - ☞ More fine-grained lexical relations
    - ☞ Minor changes to the set of semantic relations
- **Linking** the Ancient Greek and Sanskrit WNs to other textual and lexical resources
- **Applications**

# Representing meaning in ancient language WNs

- **Like previous WNs:** our ancient language WNs are lexical databases in which meaning is stored in a relational way.
- They comprise:
  1. Nodes for lemmas to which meanings are associated in the form of *synsets*;
  2. Lexical relations, which connect lemmas to each other;
  3. Semantic relations, which establish connections among *synsets*.

# Representing meaning in ancient language WNs

- Like in previous WNs:
- Our set of semantic relations **fails to capture semantic solidarity** due to belonging in the same Frame or semantic field (*tennis problem*; Fellbaum 1998: 10).

☞ No semantic relation links the AG words in 1):

- 1) *ikhthûs* 'fish'  
*thálassa* 'sea'  
*naûs* 'ship'  
*naútēs* 'sailor'  
*pléō* 'sail'

NB:

*naútēs* is morphologically derived from *naûs*, which is annotated among lexical relations.

# Representing meaning in ancient language WNs

- **Like in previous WNs:** lemmas can be assigned multiple *synsets*, which indicates **polysemy**.
- **In an innovative way: Cognitive Linguistic approach**
  - (e.g. Lakoff and Johnson 1980; Tyler and Evans 2003; Mocciaro and Short 2019)

👉 **Principled view of polysemy**, which entails:

1. avoiding exaggerating the number of distinct senses associated to a lemma;
2. assuming that all senses of a lemma can be organized in a **structured semantic network**.

# Representing meaning in ancient language WNs

- The **structured semantic network** comprises:
  1. **Literal senses**: detected based on their early attestation, concreteness, and predominance in the network (Tyler and Evans 2003: 45-50);
  2. **Non-literal senses**: derived from literal ones through the cognitive processes of **metaphor** and **metonymy**.
- Three senses associated with the adjective ***salty*** in the Princeton WordNet:
  - 2) 

<i>a. containing or filled with salt;</i>	Literal sense
<i>b. one of the four basic taste sensations; like the taste of sea water;</i>	Metonymic sense
<i>c. engagingly stimulating or provocative.</i>	Metaphoric sense

# Representing meaning in ancient language WNs

- **Salty:**

- |    |  |                  |
|----|--|------------------|
| 2) | a. <i>containing or filled with salt;</i>                                      | Literal sense    |
|    | b. <i>one of the four basic taste sensations; like the taste of sea water;</i> | Metonymic sense  |
|    | c. <i>engagingly stimulating or provocative.</i>                               | Metaphoric sense |

- Cognitive metonymy: senses associated with the polysemous word belong to the same conceptual domain;
- Cognitive metaphor: two senses belonging to different conceptual domains are mapped to one another.

# Representing meaning in ancient language WNs

- **Annotation of lemma senses:** our annotators are asked to distinguish among *synsets* that correspond to literal, metonymic, and metaphoric meanings.
- 16 *synsets* associated to the AG word ***háls* ‘salt’**, classified into three groups:

## 3) a. literal sense ‘salt’

n#05846273 | white crystalline form of especially sodium chloride used to season and preserve food

## b. metonymic sense ‘body of salty water’


n#10771040 | water containing salts

## c. metaphoric sense ‘wit’

n#05075890 | a message whose ingenuity or verbal skill or incongruity has the power to evoke laughter



# Representing meaning in ancient language WNs

- **Annotation of lemma senses:**
- **Corpus languages** that enjoy centuries of attestation and a long tradition of studies  we tag each *synsets* for:
  1. Periodization(s);
  2. Literary genre(s);
  3. (Optionally) *loci*.

Sense	Period	Genre	Loci
3)a	Archaic (8 <sup>th</sup> -6 <sup>th</sup> BCE)	poetry epic historiography narrative	Il.9.214, Od.11.123 Ar.Ach.835 Hdt.4.53
3)b	Hellenistic (323-31 BCE)	-	Call.Fr.50
3)c	Roman (31 BCE-290 CE)	philosophy treatise	Plut.2.685 Plut.2.854

Diachronic and stylistic metadata associated  
with the senses of *háis*

# Representing meaning in ancient language WNs

- **Information useful for:**

1. philologists;
2. lexical typologists;
3. historical linguists;
4. experts in neighboring fields.

- ☞ **Possible questions:**

- Do Skt., AG, and Lat. **cognate words** lexicalize comparable arrays of concepts? (etymological info)
- Do word meanings **change over time** and **vary across literary genres and authors**? *In which way?*

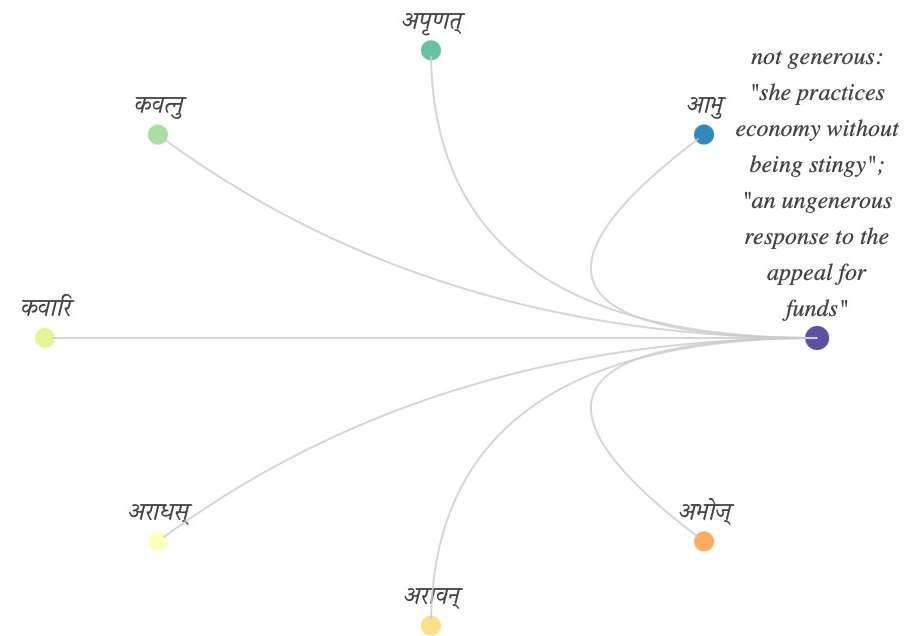
# Family-specific attributes and relations

- **Lemma attributes**

1. Etymological information
2. Rich morphological annotation

- **Lexical relations**

- **Semantic Relations**



# Lemma attributes

- **Etymological information** is hierarchically structured and consists of:

1. **ETYMOLOGY PROPER:**

PIE *\*pleu-* ‘float’ > AG *pléō* ‘sail’, Skt. *plu-* ‘float, swim’;

2. **ETYMON**, a discrete form in the history of a word’s etymological development:

PIE *\*pneu-* ‘breathe’ > AG *pneumon* ‘lung’ > AG *pleumon* > Lat. *pulmo* ;

3. **MORPHEME**, a discrete element within the etymon:

PIE *\*-ti-* in Skt. *plu-ti-* ‘flood’, AG *plú-si-s* ‘washing’.

- Each level is stored as a **separate entry** in the database ➡ lemmas can be linked via their etymological constituents at many different levels (root, stem, morphemes, etc.);
- Dedicated field with information on AG dialectal variants: Attic *plōûs* ‘sailing’, Ionic *plóos* ‘id.’

# Lemma attributes

- **Unlike in other WNs**, each lemma is provided with **rich morphological information**:

1. **MORPHO**, a ten-place character string inspired by Perseus:

AG *liměň* ‘harbor’ n-s---mn3n

2. **MORPHOLOGY**

- a. **PRINCIPAL PARTS**, where relevant parts of the paradigm are listed:

AG *háls*, Gen. *halós*;

- b. **PROSODY** providing vowel length when relevant

Lat. *occīdo* ‘to strike down’ ≠ *occĭdo* ‘to fall; die’;

3. **FORM TOKENS**: ‘irregular’ and/or ‘alternative’ forms with their morphological tag.

Form n-p---md3-, Token *hálasi*, Alternative for AG *háls*-.

# Lemma attributes

Field	Subfield	Value
Etymology	—	PIE * <i>séh<sub>2</sub></i> /- 'salt'
Lemma	—	<i>háls</i>
POS	—	Noun
Morpho	—	n-s---mn3-
Morphology	Prin. Parts	<i>halós</i>
	Prosody	—
Form Tokens	Form	n-p---md3-
	Token	<i>hálasi</i>
	Alternative	✓

Lemma annotation for AG *háls*

# Lexical relations

- In other WNs, lexical relations include:
  - **morphological relations** (derivation or composition)
  - the **semantic relation of antonymy** (speakers identify *heavy/light* as antonyms but not *weighty/light*)

! **CORPUS LANGUAGES**, no intuition of speakers:

- In our WNs, **antonymy split into a lexical** (i.e. morphological) and a **semantic relation**

☞ **Morphological antonyms** = lemma pairs, in which one of the antonyms is derived from the other through the **privative prefix *a-***:

Skt.	[ <i>a-mitra</i> - 'non-friend, enemy']	IS PRIVATIVE OF	[ <i>mitra</i> - 'friend']
	[ <i>mitra</i> - 'friend']	HAS PRIVATIVE	[ <i>a-mitra</i> - 'non-friend, enemy']

# Lexical relations

- IE languages with **rich derivational morphology** ➡ **extended the set of lexical relations:**

1. **DERIVATION**, asymmetric relation holding between a base and a word derived from it by:

- a. **conversion**      Skt.      *nāga*- A 'serpentine' > *nāga*- N 'a kind of serpent';

- b. affixation      AG      [makró-tēs ‘length’]      IS DERIVED FROM      [makrós ‘long’]  
[makrós]      IS RELATED TO      [makró-tēs]

2. **PARASYNTHESIS**, asymmetric relation holding between a base and a word derived from it by simultaneous conversion and affixation:

- AG [ánoos A 'without understanding'] IS PARASYNTHETIC OF [nóos N 'mind']  
[nóos N] HAS PARASYNTHETON [ánoos A]



# Lexical relations

## 3. COMPOSITION, asymmetric relation between a compound word and its constituents:

Skt.	[ <i>rāja-putra</i> - ‘a king’s son, prince’]	IS COMPOSED OF	[ <i>rāja</i> - ‘king’]
	[ <i>rāja-putra</i> -]	IS COMPOSED OF	[ <i>putra</i> - ‘son’]
	[ <i>rāja</i> -],[ <i>putra</i> -]	COMPOSES	[ <i>rāja-putra</i> -]

## 4. INCLUSION, asymmetric relation between a multi-word unit and its parts:

AG	[ <i>thalássia érga</i> ‘navigation’]	INCLUDES	[ <i>thalássios</i> ‘related to the sea’]
	[ <i>thalássia érga</i> ]	INCLUDES	[ <i>érgon</i> ‘work’]
	[ <i>thalássios</i> ],[ <i>érga</i> ]	IS INCLUDED IN	[ <i>thalássia érga</i> ]

## 5. PARTICIPLE, asymmetric relation between a participle and its base verb

Skt.	[ <i>sát</i> - A ‘true’]	IS PARTICIPLE OF	[ <i>as</i> - V ‘be’]
------	--------------------------	------------------	-----------------------

# Lexical relations

Relation	Label	Inverse
<b>Antonymy</b>	IS PRIVATIVE OF	HAS PRIVATIVE
<b>Derivation</b>	IS DERIVED FROM	IS RELATED TO
<b>Parasynthesis</b>	IS PARASYNTHETIC OF	HAS PARASYNTHETON
<b>Composition</b>	IS COMPOSED OF	COMPOSES
<b>Inclusion</b>	INCLUDES	IS INCLUDED IN
<b>Participle</b>	IS PARTICIPLE OF	HAS PARTICIPLE

Family-specific lexical relations

# Semantic relations

- In order to ensure **compatibility**, we try to stick to the established set as closely as possible; **but**:

1. SEMANTIC ANTONYMY, symmetric relation between *synsets*:

{n#01963712 “of moral excellence”} HAS ANTONYM {n#01078381 “having negative qualities”}

**and NOT:** AG *kalós* ‘good’ HAS ANTONYM *kakós* ‘bad’

- NB in other WNs: IS SIMILAR TO links satellite *synsets* to one of the antonyms in a cluster of adjectives;

ALSO SEE links the half cluster of adjectives to another half cluster related to it.

👉 we avoid using them!

2. IS NEAREST TO, catch-all relation for similar *synsets*:

{n#01893072 “a young pig”} IS NEAREST TO {n#01892895 “domestic swine”}

= *synsets* associated with AG *khoîros* and *sûs*

# Semantic relations

3. **VERBAL SENSE GROUP**: symmetric relation linking verbs related by aspectual, voice- or valency-related properties:

{v#00399347 “become conscious of”} VERBAL SENSE GROUP {v#00401762 “possess knowledge about”}

= *synsets* of AG *gignōskō* (PRS) ‘perceive, know’, *oîda* (PF; defective form) ‘know’

4. **QUALIFIES EVENT AS**: asymmetric relation between an adverb and an adjective:

{r#00162139 “for an extended time or at a distant time:”}

QUALIFIES EVENT AS

{a#01380813 “indicating a greater than average duration of time or a duration as specified”}

= *synsets* of AG *makrán* ‘at length’ and *makrós* ‘long’

# Semantic relations

Relation	Label	Inverse
<b>Antonymy</b>	HAS ANTONYM	HAS ANTONYM
<b>Nearest to</b>	IS NEAREST TO	IS NEAREST TO
<b>Verbal Sense Group</b>	VERBAL SENSE GROUP	VERBAL SENSE GROUP
<b>Qualify event as</b>	QUALIFIES EVENT AS	QUALIFIES ENTITY AS

Family-specific semantic relations

# Enriching the AG WordNet with sentence frames

- **PILOT STUDY:** Homeric verbs annotated in **HoDeL** (<https://hodel.unipv.it/hodel-res>)
- Ultimately, whole *Ancient Greek and Latin Dependency Treebank (AGLDT)* as well as *Vedic Treebank (VTB)*;
- **HoDeL:** corpus-based lexicon of Homeric verbs induced from AGLDT 2.0 via a series of SQL queries
  - It allows obtaining corpus-based data concerning sentence frames, their frequency, and instantiations
  - It gives information about the syntactic and semantic features of verbal dependents

# Enriching the AG WordNet with sentence frames

- Annotators are asked to extract sentence frames from HoDeL and assign them to the correct *synset*;
- Advantages of [extracting sentence frames from corpora](#):
  - 👉 Info on frequency distribution of sentence frames, the most frequent filling words for each frame-slot, and their associative connections with *synsets*
- **Case studies:**
- *pléō* 'sail, float' (28 occurrences in HoDeL)
  - 3 *synsets* 'float', 'sail', 'depart' and 2 sentence frames
- *angéllō* 'announce' (27 occurrences in HoDeL)
  - 1 *synset* 'make known' and 4 sentence frames mainly dependent on verbal aspect/tense

# Applications

- Precious resources for linguistic as well as philological and literary studies:
1. **Linguistics:** the development of objective methodologies for research on both diachronic change and interlinguistic variation in lexical semantics is a well-known *desideratum* within historical and comparative linguistics;
  2. **Philology, epigraphy, anthropology, comparative history of religions** and related fields: possibility of searching ancient texts for specific senses or semantic fields rather than for specific words (lemmas).



# Applications

- The **linking of WNs with other textual and lexical resources** reveals **connections of the semantics of lexical items** with their **syntactic contexts of occurrence** and **relevant semantic frames**;

## 1. Linguistics:

- Benefit for **Construction Grammar approaches to the study of syntax**, which assume that grammatical constructions are filled with lexical items whose meaning is coherent with that of other construction fillers and with that of the construction;
- The set of construction fillers can be easily detected by exploiting *synsets* and semantic connections between them.

## 2. Philology:

- This connection will help researchers **filling gaps in the written records**, as entire sets of near-synonymic and semantically related words will be easily made available through the WN infrastructure and user interface.

# Cylleneus

- The **linking of WNs with other textual resources** is already possible thanks to **Cylleneus**, a next-generation search engine for ancient languages (see SunoikisisDC Summer 2020, Session 10);
- It enables texts to be searched on the basis of:

- Form
- Lemma
- Morphological properties (PoS & morph)
- Syntactic properties and relations (treebanks)










→ *First-generation tools*

**But also:**

- Words' meaning
- Semantic domains
- Semantic relations

→ **Cylleneus**

# Cylleneus

	<b>Forms, wildcards</b>	'animos'	contum*
	<b>Lemmas</b>	<animus>	
	<b>Meanings</b>	[en?courage]	[it?guerra]
	<b>Semfields</b>	{611}	
	<b>Lexical &amp; semantic relations</b>	</::bellum>	[!::en?courage]
	<b>Morphology</b>	:ACC.PL.	
	<b>... as a filter</b>	[en?courage]   ABL.SG.	
	<b>Morphosyntax</b>	/ablative absolute/	
	<b>Contextual &amp; sequential queries</b>	“ . . . ”	

# Cylleneus

E.g.:

- Formulaic similes involving a cow and her calf in the *Ṛgveda*;
- Schematic formula: VERB [COW] *ná/iva* 'like' [CALF]
- **Query** [en?cow] \ [en?calf]

- *dhenú-* 'milch cow' + *vatsá-* 'calf'

# dhenúr ná vatsám | páyasā abhí 10 11 12 #

# dhenúr ná vatsám | yávasasya pipyúṣī #

# 1 2 vatsám ná | svásareṣu dhenáva #

# 1 2 vatsám ná | svásareṣu dhenáva #

# abhí vatsám ná dhenávaḥ #

- *mātr̥-* 'mother' + *śíśu-* 'baby'

# 1 2 3 4 5 | śíśum ín ná mātáro #

- *mātr̥-* 'mother' + *vatsá-* 'calf'

# 1 2 vatsám ná mātáraḥ #

# 1 2 vatsám ná mātáraḥ #

# gāvo vatsám ná mātáraḥ #

- *vāśrá-* 'bellowing' *dúhāná* 'milker' + *vatsá-* 'calf'

# vāśréva vatsám | 6 7 8 dúhānā #

- ...