Multilingual data in ELTeC: enacting European literary traditions

Mini-workshop

Roxana Patraș, Ioana Alexandra Lionte

Workshop structure and goals

1. Greetings and getting to know each other: brief introduction of workshop creators
2. An introduction to the Distant Reading for European Literary History project and a subsequent description of ELTeC (European Literary Text Collection)
3. A discussion about the multilingual challenges entailed by ELTeC structure and sampling principles
4. An exploration of the ELTeC titles
5. An interactive outlook on a possible frame of annotation

Distant Reading for European Literary History and the European Literary Text Collection (ELTeC)

- *ELTeC structure and collections*
- *sampling principles*
- *level 0 and 1 encoding*
- *linguistic traditions and literary typologies: comparative challenges*

ELTeC

- EU-funded COST Action “Distant Reading for European Literary History” (CA 16204)
- European network with 35 members from 23 countries
- An open source multilingual benchmark corpus for European literature (ELTeC)
- Focus on the novel

Why focus on the novel?

Students, write your response!

Which of the following criteria should be taken into consideration when focusing on the novel?

ELTeC in numbers

4 Working Groups

15 European languages

100 novels per country

1840-1920

Language	Texts	Words	Male Author	Female Author	Short	Medium	Long	1840-59	1860-79	1880-99	1900-20	Frequent	Rare
cze	16	366626	14	2	16	0	0	5	6	5	0	0	15
deu	98	12086096	65	33	20	37	41	24	24	25	25	46	46
eng	100	11907427	50	50	26	29	45	22	23	29	26	28	44
fra	100	7990954	64	36	30	43	27	25	25	25	25	35	42
gre	11	42524	10	1	11	0	0	0	1	6	4	3	4
hun	100	7591321	85	15	44	33	23	24	24	25	27	41	31
ita	34	3328244	32	2	13	10	11	5	12	10	7	12	0
lit	17	497663	13	4	13	3	1	5	3	4	5	5	12
nor	27	1114092	22	5	18	9	0	2	2	19	4	26	1
por	98	6565452	82	16	42	38	18	12	32	19	35	22	22
rom	66	3901964	55	7	31	25	10	3	14	20	29	23	43
slv	100	5682120	89	11	53	39	8	2	13	36	49	48	52
spa	44	3908459	32	12	12	21	11	10	14	13	7	28	16
srp	45	1723848	38	7	32	13	0	0	5	18	22	14	25

Last updated: 2020-05-01

Sampling criteria

- Language: European languages, no translations
- Prose: narrative fictional prose
- Period: 1840-1920
- Length: min. 10.000 words
- Publication: a preference for books over novels published in serial publications
- Access: only freely available digitizations

Which of the following sampling criteria do you see fit for the purpose of ELTeC?

I will not define what a novel is.

I follow a non-normative but metadata-based approach (not canon-based).

I aim to represent the variety of a population.

I allow for a comparability of texts and individual sub-collections according to different metadata set(s).

Students, draw anywhere on this slide!

Guidelines for corpus building

1. Author gender: female, male, mixed/diverse/undefined. Novels by female authors need to make up at least 10% of the novels in each collection.
2. Length (words): short (10-50k words), medium (50- 100k words) and long (more than 100k words). At least 20% of the novels should be short and 20% should be long.
3. Publication time: **1840-1859**, **1860-1879**, **1880-1899** and **1900-1920**. Each group of novels should make up approximately one quarter of the collection.
4. Reprint count of each novel, specifically in the period 1970-2010, as an operationalization of one aspect of canonicity, in two groups: low (no reprint) and high (one or more reprints).

TEI teiHeader profileDesc

```
<encodingDesc n="eltec-1">
        <p/>
</encodingDesc>
<profileDesc>
 <langUsage>
  <language ident="ro"/>
 </langUsage>

 <textDesc>
  <e:authorGender key="M"/>
  <e:size key="medium"/>
  <e:canonicity key="high"/>
  <e:timeSlot key="T2"/>
 </textDesc>

</profileDesc>
<revisionDesc>
 <change when="2019-11-15">Initial TEI version generated.</change>
</revisionDesc>
</teiHeader>
<text xml:lang="ro">
<front>
 <div type="titlePage">
     <p>CLASICII</p>
     <p>ROMÂNI</p>
     <p>COMENTAȚI</p>
         <milestone unit="line"/>
     <p>SUB ÎNGRIJIREA</p>
     <p>DLUI N.CARTOJAN</p>
     <p>PROFESOR UNIVERSITAR</p>
         <milestone unit="line"/>
     <p>N.FILIMON</p>
     <p>CIOCOII</p>
     <p>VECHI ȘI NOI</p>
```

Encoding. TEI (Text Encoding Initiative)

- Encoding standard and guidelines for the representation of texts for humanities
- For various texts, e.g. manuscripts and prints, books, letters, poems, and dictionaries:
  - ➢ Text-internal categories, text-external categories
  - ➢ Mark up, text structure(s) and divisions, content and references
- Guidelines provide ca. 500 elements and various specifying attributes

Encoding Requirements

The ELTeC encoding scheme was deliberately not intended to represent source documents in all their original complexity of structure or appearance, but rather to make it as simple as possible to access the words of which texts are composed in an informed and predictable way.

How can this be achieved?

Students, write your response!

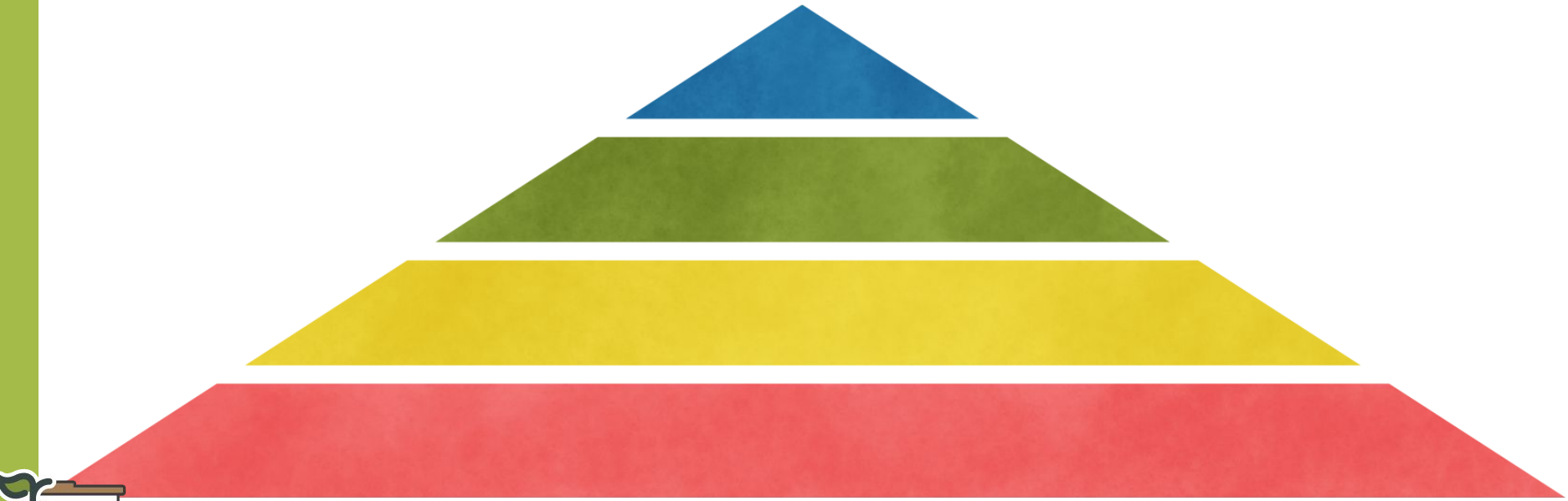
An important principle following from this latter goal is that ELTeC markup should offer the encoder very little choice, and the software developer very few surprises: the number of tags available is greatly reduced, and their application is tightly constrained.

ANNOTATION

- explicit assignment of categories to one or more exponents in a corpus, always interpretation (c.f. a.o. Lüdeling 2011; McEnery and Hardie 2012; Zinsmeister et al. 2008)

- tag set and guidelines: defining categories (and values) and formulate guidelines on how and when to assign them

What are the steps encountered so far? (sampling/guidelines in corpus creation, encoding requirements, annotation)



Students, draw anywhere on this slide!

ELTeC encoding scheme/s

level 0: minimal encoding scheme for texts produced manually or by OCR from print originals

level 1 : somewhat richer format derivable automatically from texts encoded in other formats (Word, HTML TEI ...)

level 2 : linguistically annotated and segmented and a tightly constrained Header common to each level

Level 0

discard non authorial front or back matter

distinguish titlepage from other front matter mark chapter divisions and headings, but no substructure

mark paragraphs (MLE blocks of text)

reassemble words broken across lines

discard paratext, illustrations, notes, corrections

(optionally) mark pagebreaks and highlighting

text may or may not be normalized: no indication either way

Never did she appear to more advantage, for although her dress was only white muslin, enlivened by a gold band round her waist, it fitted exquisitely, displaying her beautiful figure to the fullest perfection, and her simple *coiffure*, glossy luxuriant hair, unencumbered by flowers, or any of the superfluous ornaments with which young ladies *will* disfigure themselves,

allowed the beholder to feast his eyes upon the statue-like shape and proportion of the small undecorated head.

And Car was at that moment thoroughly pleased, and that alone gave an additional charm to her face.

```
<div type="chapter" n="23">
<head>Chapter XXIII.</head>
<!-- ... -->
<p> Never did she appear to more advantage,
for although her dress was only white muslin,
enlivened by a gold band round her waist, it
fitted exquisitely, displaying her beautiful
figure to the fullest perfection, and her
simple <foreign>coiffure</foreign>, glossy
luxuriant hair, unencumbered by flowers, or
any of the superfluous ornaments with which
young ladies <emph>will</emph> disfigure
themselves, <pb n="302"/> allowed the beholder
to feast his eyes upon the statue-like
shape and proportion of the small undecorated
head. </p>
<p> And Car was at that moment thoroughly
pleased, and that alone gave an additional
charm to her face.</p>
<!-- ... -->
</div>
```

Level 1

mark chapter substructure with <milestone> and <label>

elements

interpret highlighting, where possible, using <emph>

<foreign> or <title>

record authorial notes (gathered together into back)

record graphics as <gap>

normalized forms are explicit but original forms are lost

```
<text xml:lang="ro">
<front>
 <div type="titlePage">
     <p>CLASICII</p>
     <p>ROMÂNI</p>
     <p>COMENTAȚI</p>
         <milestone unit="line"/>
     <p>SUB ÎNGRIJIREA</p>
     <p>DLUI N.CARTOJAN</p>
     <p>PROFESOR UNIVERSITAR</p>
         <milestone unit="line"/>
     <p>N.FILIMON</p>
     <p>CIOCOII</p>
     <p>VECHI ŞI NOI</p>
     <p>ROMAN SOCIAL</p>
      <p>COMENTAT DE:</p>
  <p>GEORGE BAICULESCU</p>
  <p>BIBLIOTECAR LA ACADEMIA ROMANA</p>
  <p>EDITURA „SCRISUL ROMÂNESC”/CRAIOVA</p>
 </div>
  <div type="liminal">
      <div>
      <p>CIOCOII VECHI SI NOI</p>
      <p>SAU „CE NAŞTE DIN PISICĂ</p>
      <p>ŞOARECI MĂNÂNCĂ"!</p>
      <milestone unit="line"/>
   <p>DEDICAŢIE</p>
      <milestone unit="line"/>
  <p><foreign><hi>Domnilor Ciocoi!</hi></foreign></p>
  <p>Este mult timp de când umblu cu această nuvelă ziua şi noaptea, întocmai ca Diogen, căutând o clasă de oameni ca să le-o dedic. Am voit să fac această onoare
boierilor; dar, după o gândire serioasă, mi-am schimbat hotărîrea, căci de şi într’această clasă s'au strecurat mulţi venetici conrupţi şi cu toate lovirile şi
tentaţiunile străinilor la cari servă de ţintă de un secol şi jumătate, tot se găsesc printre dânşii bărbaţi cu simţiminte nobile şi cu inimă de adevăraţi Români, cari
au făcut, fac, şi sunt convins că vor face, mult bine patriei lor.</p>
```

Go to definition of the 'div' element

N. FIL'I'MON

CIOCOII

VECHI SI NOI

sl ROMAN ¿OCIAL

DUiCA

Id

CIOCOII VECHI ŞI NOI

SAU „CE NAŞTE DIN PISICĂ ŞOARECI MĂNÂNCĂ"!

DEDICAŢIE

Domnilor Ciocoi!

Este mult timp de când umblu cu această nuvelă ziua şi noaptea, întocmai ca Diogen, căutând o clasă de oameni ca să le-o dedic. Am voit să fac această onoare boierilor; dar, după o gândire serioasă, mi-am schimbat hotărîrea, căci de şi în-' tr'această clasă sau strecurat mulţi venetici con-rupţi şi cu toate lovirile şi tentaţiunile străinilor la cari servă de ţintă de un secol şi jumătate, tot se găsesc printre dânşii bărbaţi cu simţiminfe nobile şi cu inimă de adevăraţi Români, cari au făcut, fac, şi sunt convins că vor face, mult bine patriei lor.

Dela boieri am alergat la negustori. Am revizuit toate stabilimentele de comerţ, dela magaziile cele mari şi luxoase până la. magherniţele cele umilite ale precupeţilor. Am văzut zarafi fără capital, fanfaroni şi malonești[1]) cari sărăcesc lu> mea prin dobânzile lor cele nemăsurate; lipscani

şi bogasieri[2]) cari îşi împodobesc magazinele cu marfă putredă şi cu oglinzi mincinoase şi, dân-du-şi ton de mari capitalişti, ruinează societatea prin falimente frauduloase, ce se efectúese foarte lesne în ţara noastră; băcani cari vând rapiţă în loc de untdelemn, orez îndoit cu pietricele ca să tragă mai greu la cântar şi cafea amestecată cu orz şi fasole. Am văzut cârciumari amestecând vinul cu apă şi vânzând cu ocale cu două funduri, măcelari şi precupeţi vânzând cu cântare strâmbe, şi m'am mâhnit, căci rău este foarte mare, dar n'am găsit în aceşti amăgitori decât nişte hoţi sau ciocoiaşi ordinari, ieşiţi din şcoala voastră fără diplomă de specialitate!... Am alergat prin sate şi cătune, am vorbit cu ţărani bătrâni şi tineri; ce e drept, sunt plini şi ei, sărmanii, de mârşevii până între urechi, dar n'am găsit nici între dânşii pe oamenii ce căutam. Am intrat în locaşul lui Dumnezeu, am observat cu conştiinţă clerul înnalt şi pe cel proletar. Dar vai! ce dezamăgire!... Acolo unde credeam că voi găsi toiagul şi traista, sacrul simbol al umilinţii şi pietăţii creştine, am găsit: ignoranţa întronată, invidia, mândria lăcomia şi alte păcate mortale, pe care ne oprim a le descrie, căci legea de presă, fără îndoială, ne-ar condamnă la zece ani de ocnă.

Challenges

- creating a single abstract model of textual components, which might usefully be considered independently of its expression in a particular source or output
- cultivating mutual respect for the widely differing scientific, cultural, and linguistic traditions characterising this cross-European and cross-disciplinary project
- the design of an encoding scheme that should not rely on any pre-existing notion of textual ontology
- creating a principled and representative selection.

What should be taken into account when creating a representative selection? :

Population

Culture

Gender

Heritage

Canon

Representativeness

Yes

No

Maybe

Students, draw anywhere on this slide!

- Lack of data about the population we are claiming to represent -- which is hard to come by for many of the languages concerned

- The novels which we know about tend to be the ones that national libraries or equivalent cultural heritage institutions have chosen to preserve, which publishers over time have been able to sell, and which lecturers in literary studies have chosen to teach.

- Discarded or even suppressed titles as unworthy of inclusion in the national patrimony

But how can we express opinions about changes in the nature of the published novel if the sample on which we base those opinions is wildly different in composition from the actual population?

If our data leads us to assert that novels in a given language are never written by women, or are never of fewer than 100,000 words is this simply because no female authors happen to have been preserved, or because short novels were routinely discarded from the collection?

On the other hand, does this actually indicate something fundamental, a characteristic of the population we are investigating?

Students, draw anywhere on this slide!

Remember the guidelines of corpus building:

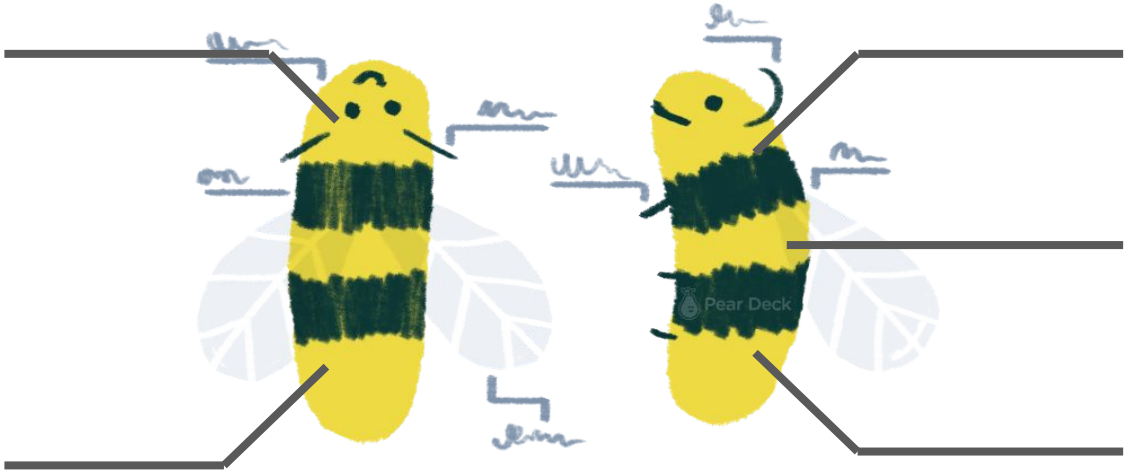
Word Bank

Author gender

Length

Publication time

Reprint count



Students, draw anywhere on this slide!

Country-specific challenges

The French ELTeC

- multiplicity of different platforms
- limited amounts of useful metadata that is available
- none of the available catalogues features information on more than one of our two eligibility criteria (length, period) and our four key corpus composition criteria (YP, AG, S, RC)
- multiplicity of different platforms that offer digital texts and the many different formats that these platforms have

The Portuguese ELTeC

- full-text availability of public-domain Portuguese literature
- the absence of reliable metadata for most collections
- Portuguese literature and texts span many centuries and genres
- digitalization efforts end at the pdf level, and have often been focused on important (canonical) authors, or on thematic issues

The Romanian ELTeC

- underdeveloped library services and librarians' training
- scarce open access resources and available formats
- OCR and POS tagging sub-optimal performance on diachronic varieties of Romanian
- data on booklength available in page numbers but not in word count
- unbalance of T1, T2, T3, T4; the percentage of female-authored novels
- the Romanian team's low training level and lack of experience with xml, epub, hml formats

Are you enjoying this workshop?

An exploration of ELTeC titles

- *presentation of the project on ELTeC titles*
- *presentation of the collections chosen for annotation*
- *discussion of annotation guidelines*
- *interactive case study*

ELTeC titles

- title as *predictive instruments/ pointers* for how "the great unread" in the ELTeC corpus/ data could be reclusterd in order to reassess the traditional assumptions on "canonization", "genre", "periodization", etc.

- two different ways of preparing data for analysis: manual annotation processing via UDPipe (POS tagging)

Ways of looking at ELTeC titles

Predictive instruments for reclustering the “great unread” in the ELTeC corpus

User-oriented, changing in time according to conventions, the most mobile part of the entire book

Threshold that blurs the boundaries between the story (fiction) and the real world (fact)
G. Genette

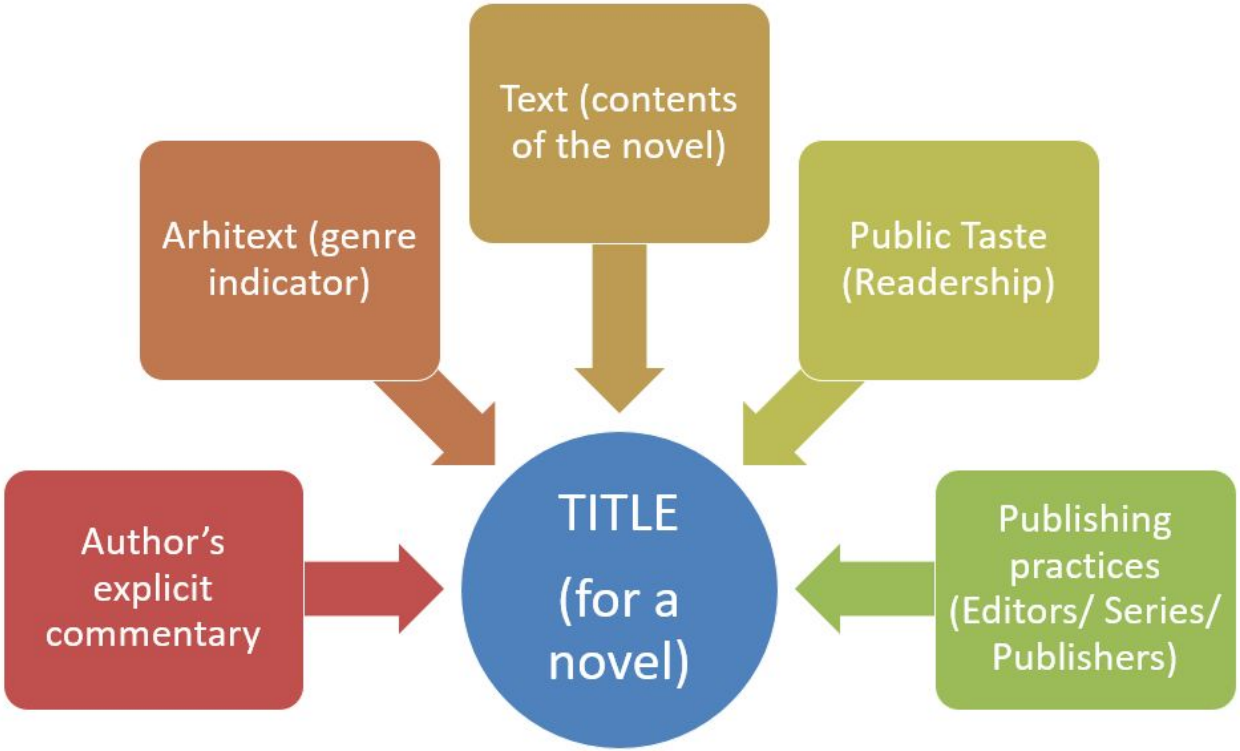
Complex whole, container and content at the same time
Levin

Basic structuring principle of the literary narrative
A. Jolles

The only explicit commentary that the reader is given from the author
Booth

Students, draw anywhere on this slide!

A title's referential world



11 collections: DEU, ENG, ITA, POL, FRA, POR, ROM, SPA, SLV, SRP, UKR

Students, draw anywhere on this slide!

Annotation guidelines

- POS tagging
- each title treated as a “sentence”
- titles as POS sequences
- POS distribution and relevance (nouns/nominal groups/verbs)

Circle how you are feeling:

Students, draw anywhere on this slide!

Interactive activity

Step 1: You will receive 11 Excel documents containing the ELTeC titles from 11 countries.

Step 2: Look at those written in the languages you are familiar with for a couple of minutes and notice the diversity/recurrence in their structure.

Step 3: Take your time in order to do the following activities.

1. What guidelines of title annotation would you propose in order to create a cross-linguistic schema rather than a language-specific one?

2. Which categories for semantic annotation would you create by looking at the titles?

What subcategories would you propose for the “person” category?

Word Bank

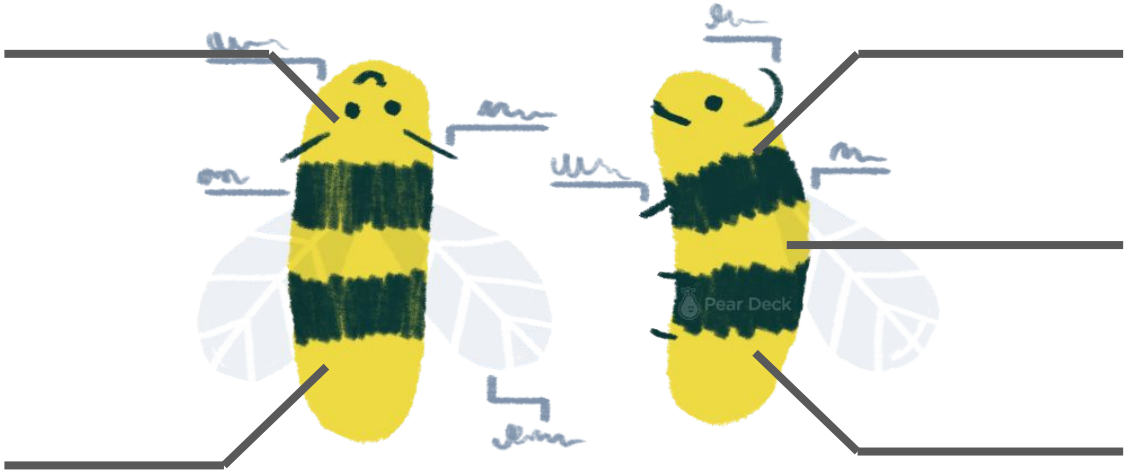
Word 1

Word 2

Word 3

Word 4

Word 5



Students, draw anywhere on this slide!

What subcategories would you propose for the “place” category?

Word Bank

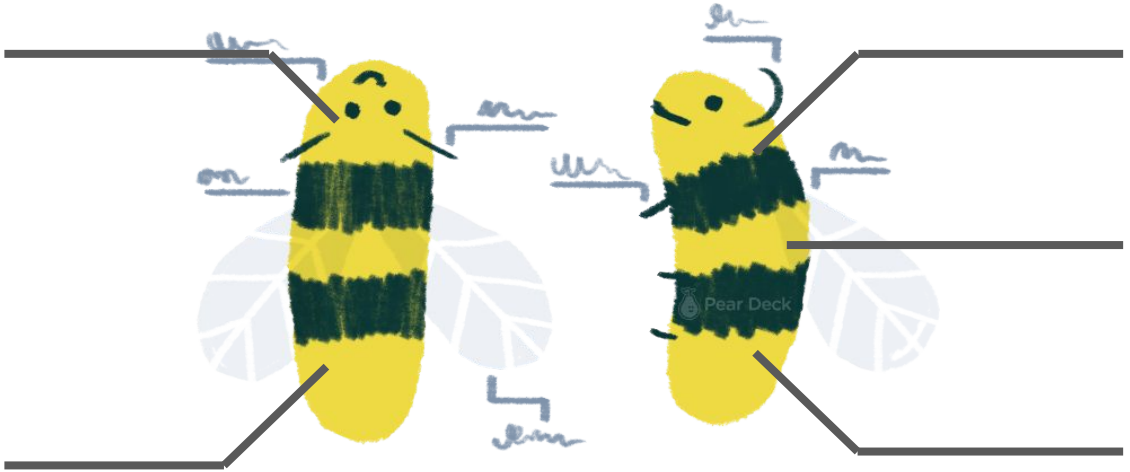
Word 1

Word 2

Word 3

Word 4

Word 5



Students, draw anywhere on this slide!

Categories for semantic annotation

- PLACES: entity, attribution, determiner
- PERSONS: status, entity, attribution, determiner/article, gender
- OTHER: entity, attribution, determiner/article
- STRUCTURE: structure complexity (no. of elements), marker of complexity
- GENRE POINTERS: how many titles indicate genre?

Full annotation see https://github.com/distantreading/WG1/blob/master/titlePilotStudy/data/dataPreparation.md

In one minute, write the thing you liked most about today's workshop.

Students, write your response!

Thank you for your attention!

Special thanks to Ioana Galleron and Carolin Odebrecht whose study *Titles in ELTeC: “thresholds” to “the great unread”* was instrumental in the creation of this workshop.

Sources:

- Carolin Odebrecht, Christof Schöch, Lou Burnard, Borja Navarro-Colorado et al. (2019), PARTHENOS Workshop for CEE countries, *COST Action Distant Reading for European Literary History Corpus Design Principles and Challenges.*
- Lou Burnard, Christof Schöch, Carolin Odebrecht, *In search of comity: TEI for distant reading*.
- Roxana Patraș, Ioana Galleron, Carolin Odebrecht, Titles in ELTeC: “thresholds” to “the great unread”. An exploration of the ELTeC paratext.

Links:

https://distantreading.github.io/ELTeC/index.html

https://www.distant-reading.net/

https://github.com/distantreading/WG1/tree/master/titlePilotStudy

https://proiectulbrancusihairo.wordpress.com/