Titles in ELTeC: thresholds to "the great unread"

An exploration of the ELTeC paratext

Roxana Patras, Ioana Galleron, Carolin Odebrecht

Goal of this workshop contribution

- Incentive for all WG members to either contribute to this study or to start new ones
- Simple work in progress example for starting a discussion on what we could do with ELTeC

Plan

- Explorative pilot study on the ELTeC paratext
- 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.

- Introduce two different ways of preparing data for analysis
- Manual annotation
- Processing via UDPipe (POS tagging)

- explore data and discuss methodological issues
- Distinction among "title", "secondary title", "subtitle" in the TEI header.
- Correlation between data extracted from titles and ELTeC body texts.

Ways of looking at ELTeC titles:
Why are the ELTeC titles different from titles listed in a library or editor catalogue, index, dictionary, bibliography?

- ***title*** as predictive instruments for how "the great unread" in the ELTeC corpus/ data could be reclustered
- ***title*** as metadata vs. "estrangement" device that seems most characteristic to DR approach ("distance as a condition of knowledge"): user-oriented, changing in time according to conventions, the most mobile part of the entire book
- ***title*** as "threshold" (G. Genette 1987) that blurs the boundaries between the story (fiction) and the real world (fact), that tests the boundaries of genre (Encyclopedia of the Novel 2011), that mediates between familiarity and defamiliarization (estrangement effect), between close reading and distant reading

Ways of looking at ELTeC titles:
Why are the ELTeC titles different from titles listed in a library or editor catalogue, index, dictionary, bibliography?

- ***title*** not as element of a whole but rather as "complex whole" itself, whose complexity is not exactly due to length but to its "underlying duality", as "container" and "content" at the same time (Levin 1977)
- ***title*** as "simple form" (*riddle*, *memorabilia*?), thus as a basic structuring principle of the literary narrative (A. Jolles 1929)
- ***title*** as the only "explicit commentary" that the reader is given from the author (Booth 1961)

Preliminary information

- 6 collections analysed: FRA, POR, ITA, ROM, ENG, SPA
  - as far as there were updated in January 2020
- Different numbers of titles: 81, 69, 34, 31, 94, 15 (total 324)
- The titles in 3 collections have been checked with original (first print titles): ROM, ITA, SPA
  - Thus, there might be discrepancies between the titles of first editions and later editions that needs further checking.
- Total number of tokens (word forms and punctuation): 1614

Methodology: POS tagging with UDPipe

- Each title treated as "a sentence"
- Post-processing:

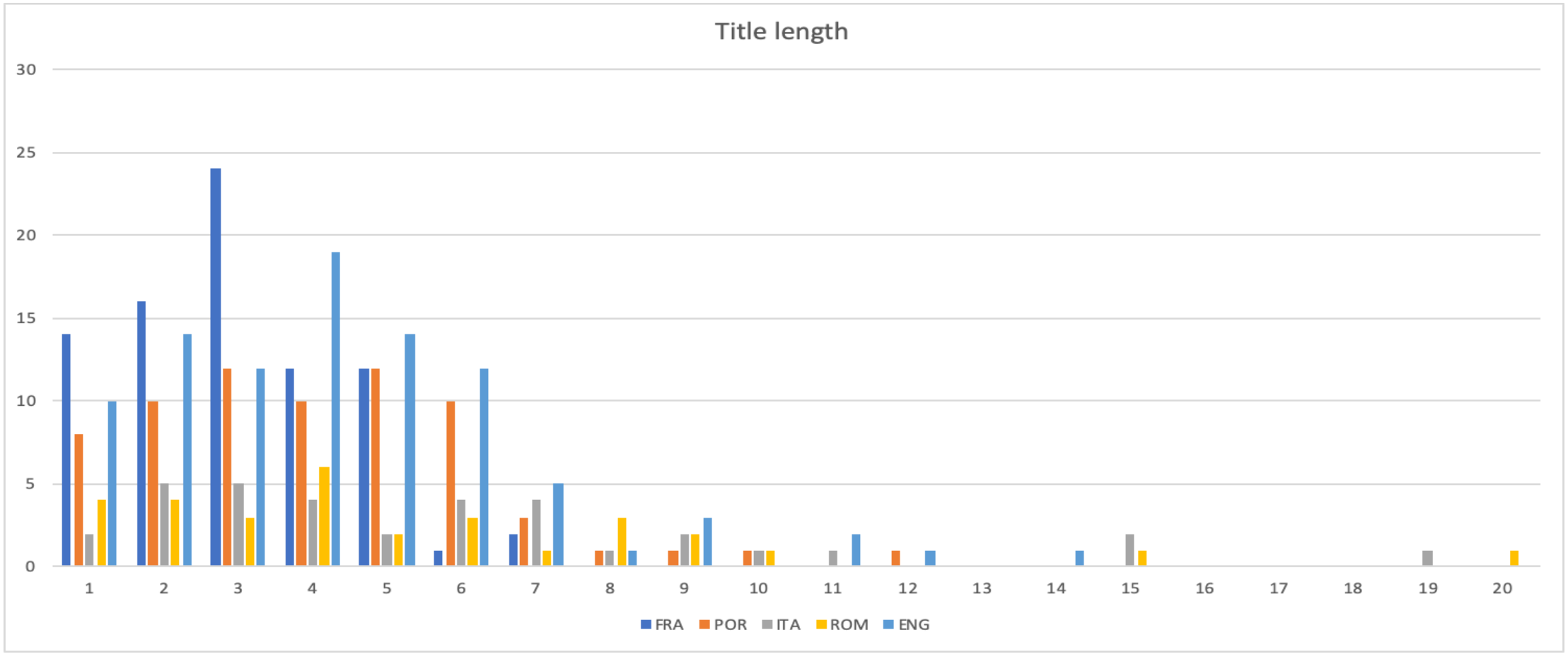
  - putting together the two-element titles that UDPipe treats as separate sentences (e.g. *Profumo. Romanzo* --> "*profumo romanzo"*)

  - correcting the POS tagging
- Titles as sequences of POS (e.g. "*under western eyes"* > ADP ADJ NOUN)

Limitation to this pilot study

- The following slides show some explorative data visualisations which show frequencies of categories attached to titles.
- The data reflects a subset to the current version of ELTeC. The different amount of titles is due to the different growing of the language collection of ELTeC.
- As ELTeC's corpus design requires to represent a certain variaty, e.g. in terms of publication date, we cannot draw any conclusions to diachronic changes.
- We therefore do not show any firm results and cannot draw any conclusions. However, it is possible to use these plots to unravel ideas and approaches for research questions and projects.

Title lengths

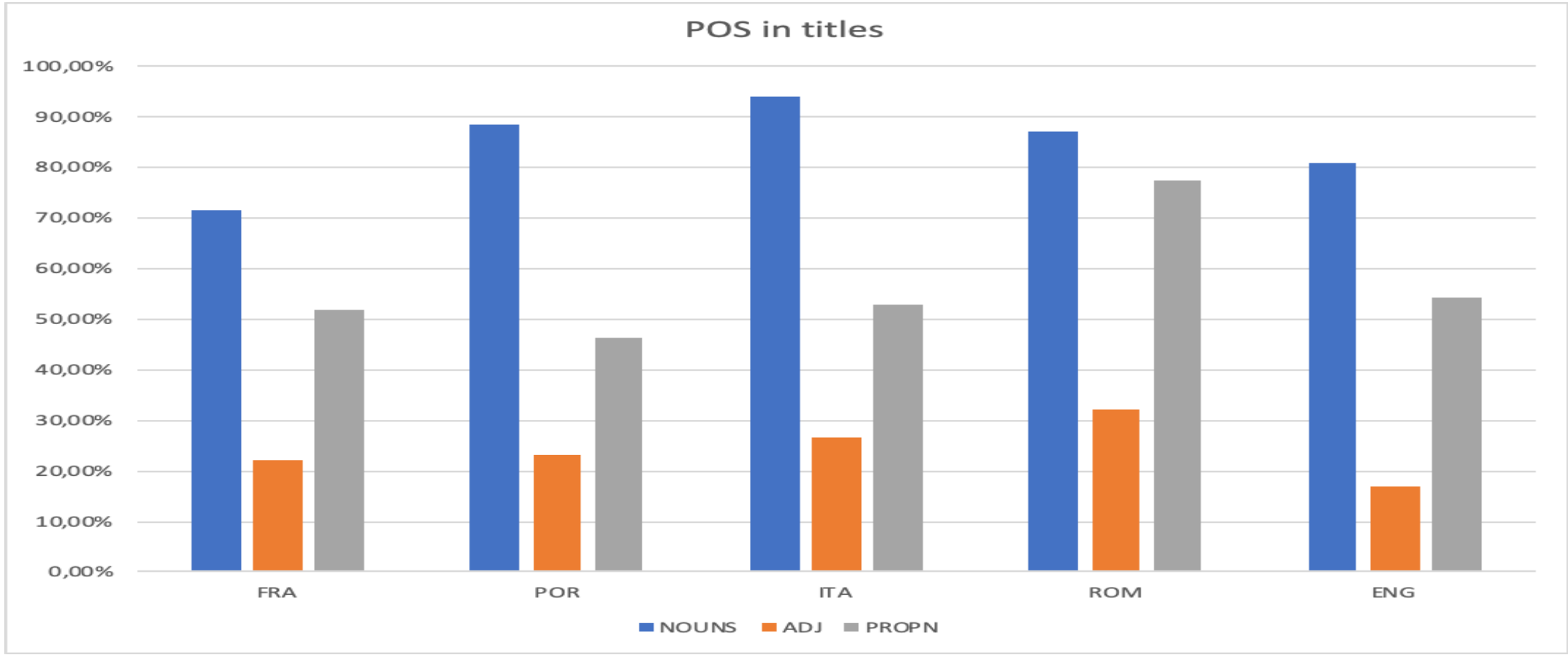


y=count of titles, x=counts of tokens within titles

Title composition (POS distribution)

- Obviously, nouns are the most frequent POS in titles, since they are an excellent tool for topicalisation > it would be interesting to look at verbs (where and when do they appear?)
- Nominal groups are to be analyzed as simple or as expanded? >>> we looked at the adjectives as the most obvious expansion of the noun
- What about proper names?

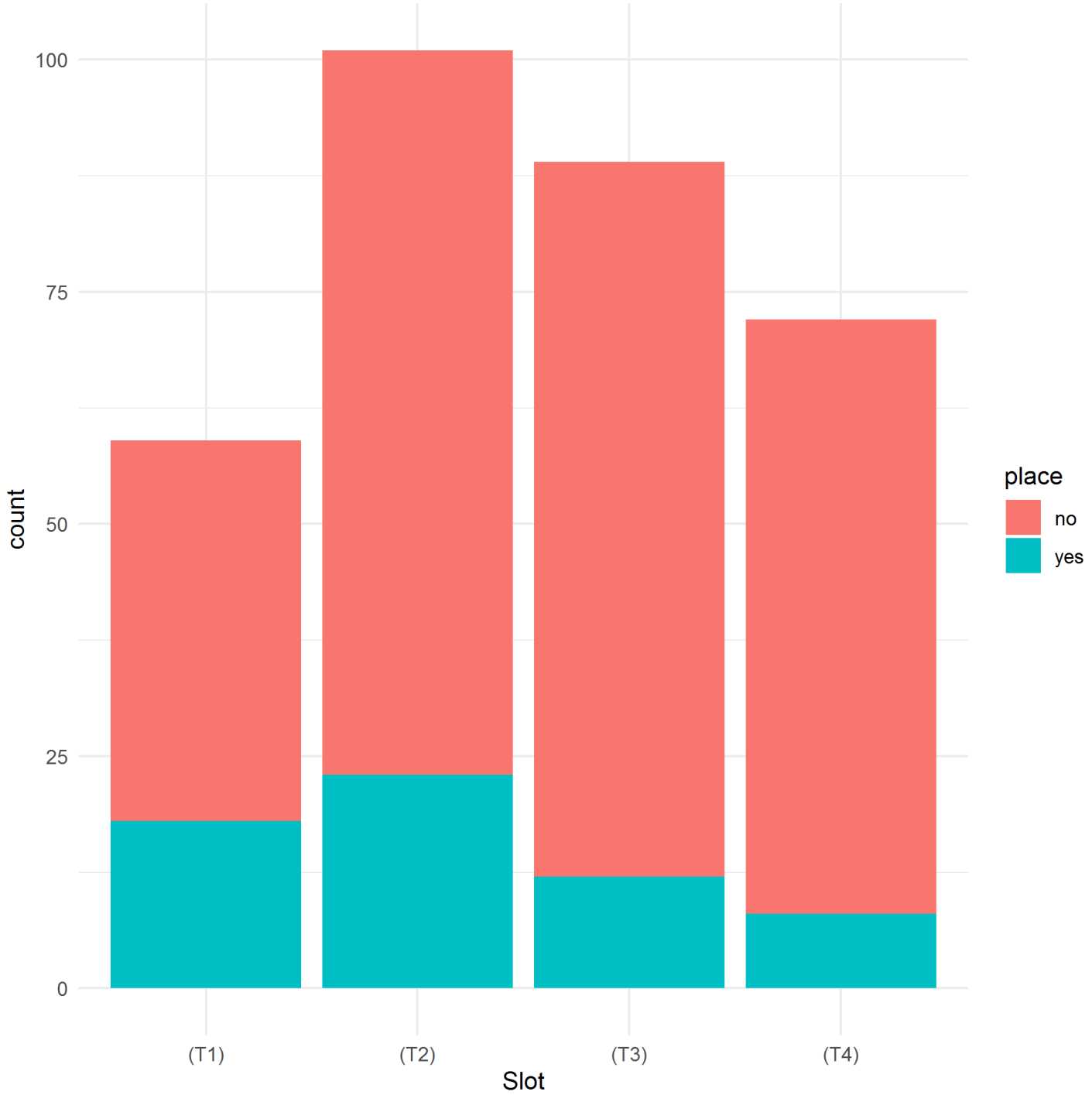
Title composition

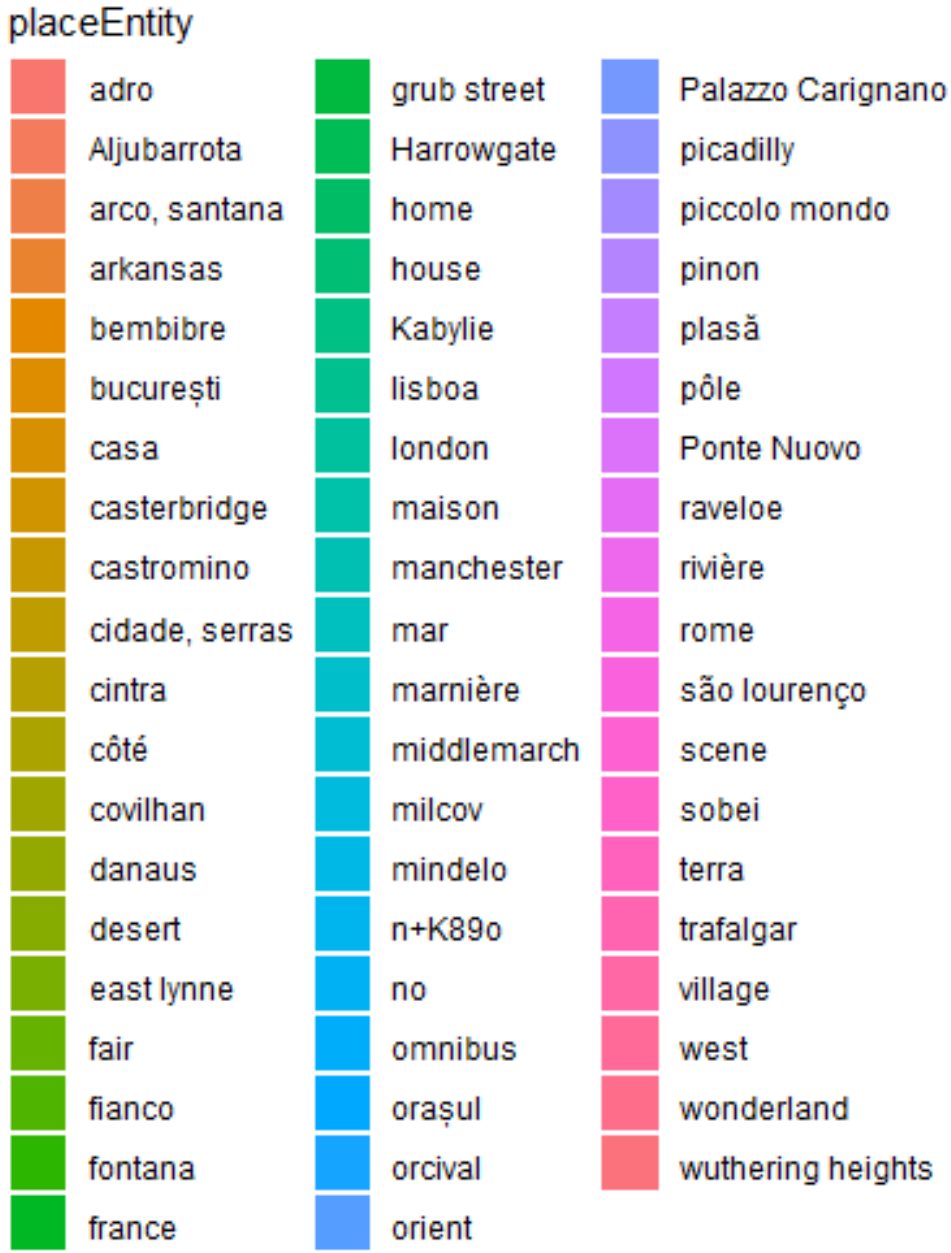


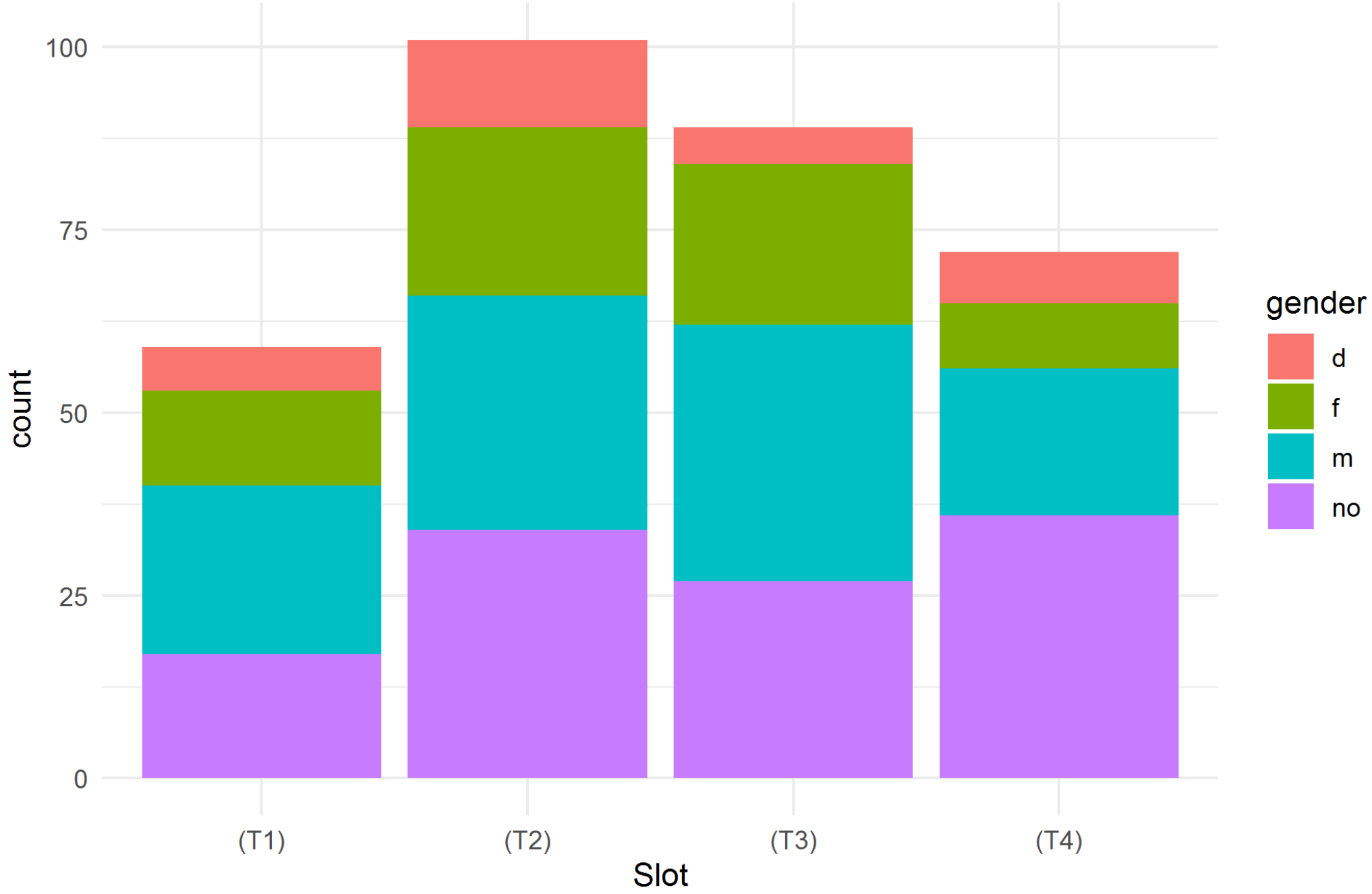
Title references

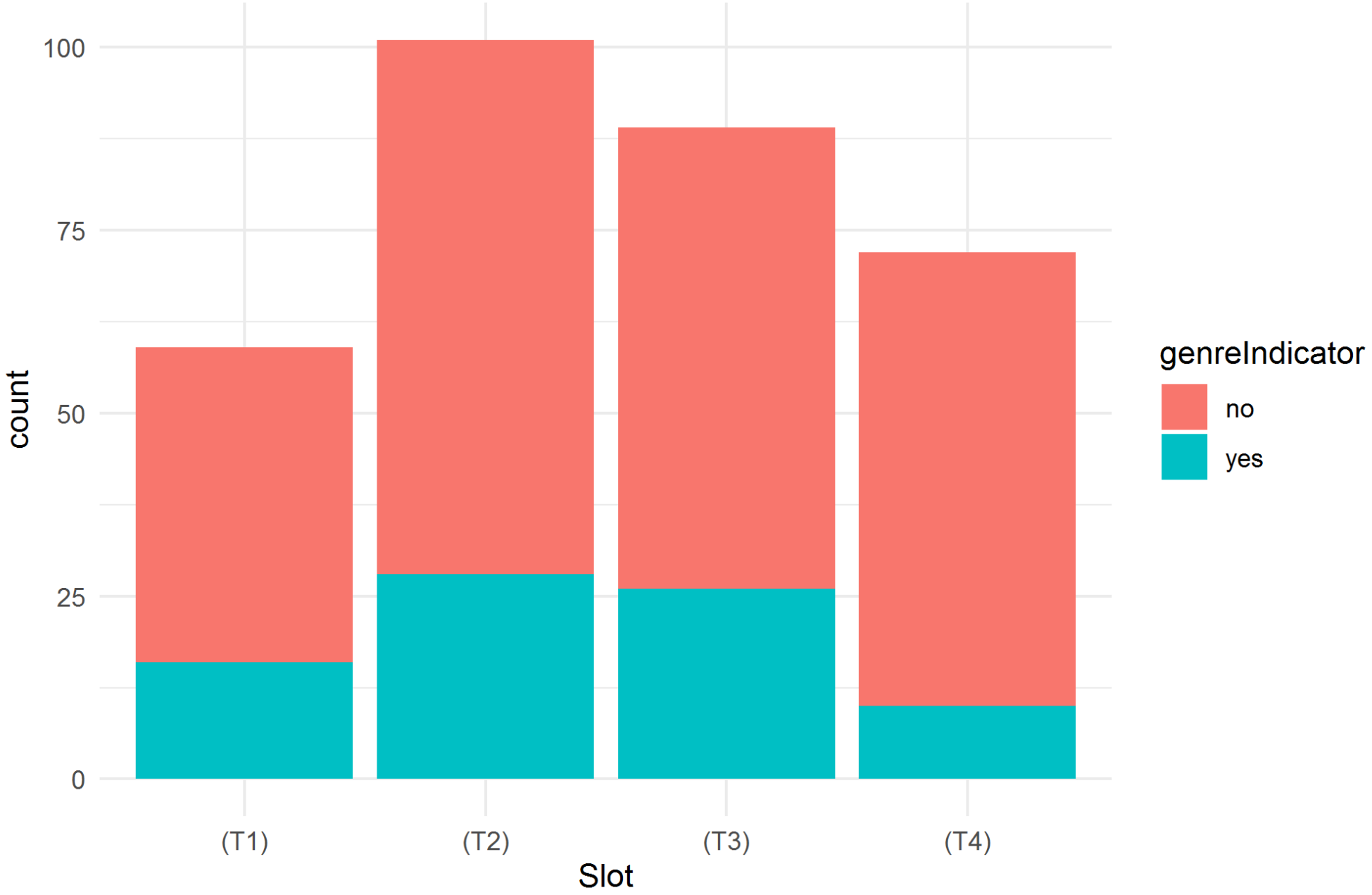
1. places
2. person, gender and status as attributes to persons
3. structure
4. genre indicator

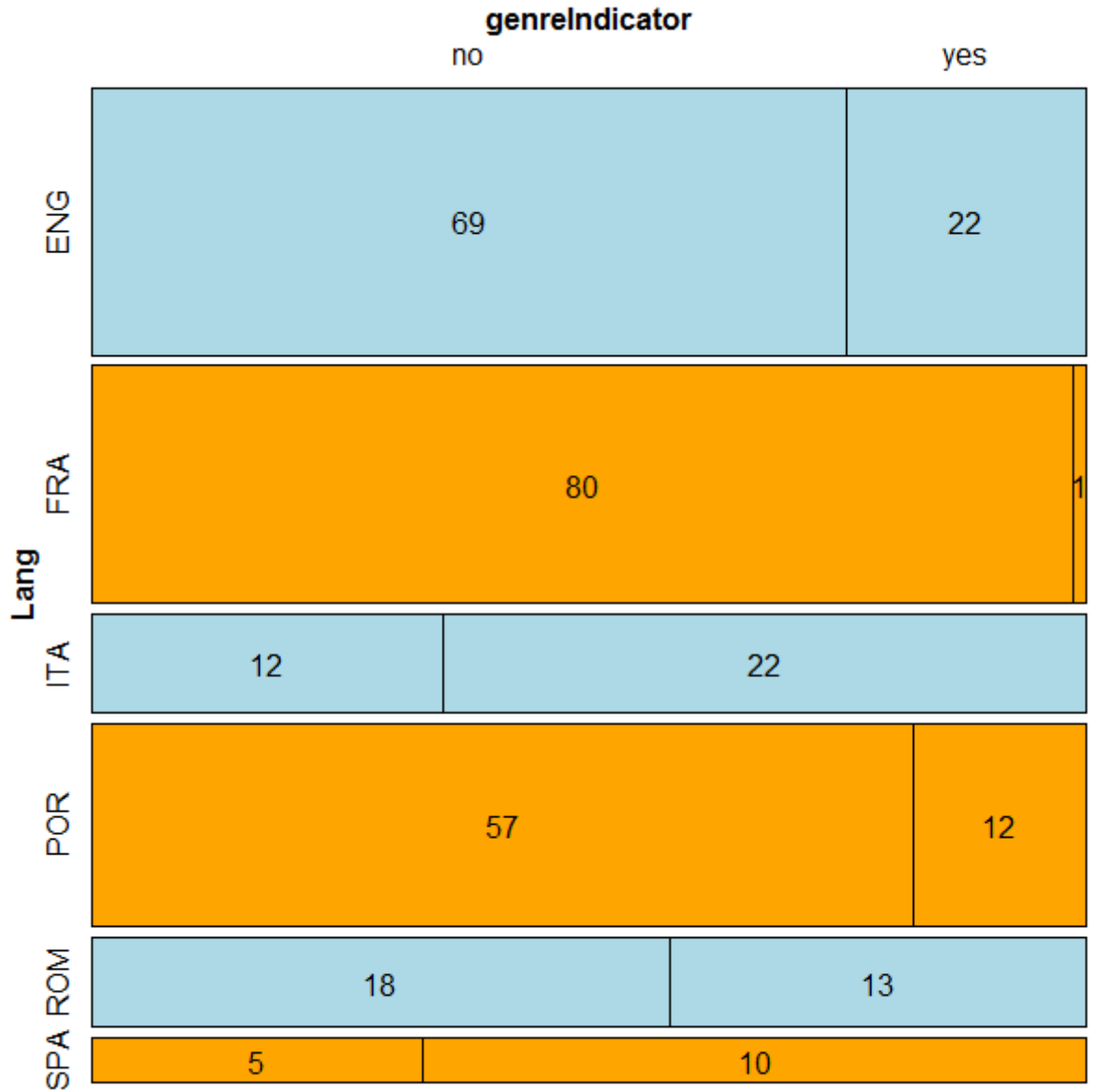
Full annotation see https://github.com/distantreading/WG1/blob/master/titlePilotStudy/data/dataPreparation.md


100
75
50
25
0
count
(T1)
(T2)
(T3)
(T4)
Slot
place
no
yes


placeEntity
adro
Aljubarrota
arco, santana
arkansas
bembibre
bucureşti
casa
casterbridge
castromino
cidade, serras
cintra
côté
covilhan
danaus
desert
east lynne
fair
fianco
fontana
france
grub street
Harrowgate
home
house
Kabylie
lisboa
london
maison
manchester
mar
marnière
middlemarch
milcov
mindelo
n+K89o
no
omnibus
oraşul
orcival
orient
Palazzo Carignano
picadilly
piccolo mondo
pinon
plasă
pôle
Ponte Nuovo
raveloe
rivière
rome
são lourenço
scene
sobei
terra
trafalgar
village
west
wonderland
wuthering heights


100
75
50
25
0
count
(T1)
(T2)
(T3)
(T4)
Slot
gender
d
f
m
no


100
75
50
25
0
count
(T1)
(T2)
(T3)
(T4)
Slot
genreIndicator
no
yes


genreIndicator
no
yes
Lang
ENG
69
22
FRA
80
1
ITA
12
22
POR
57
12
ROM
18
13
SPA
5
10

ENG	FRA	POR	ITA	SPA	ROM
novel tale novel tale story biography story history study sketch romance novel romance autobiography novel novel diary story story novel episode story	roman	episodios romance romance estudo conto episodio romance romance- cronica narrativa romance diario escriptos	avventure storia romanzo libro romanzo storielle romanzo romanzo romanzo romanzo racconto romanzo romanzo romanzo romanzo romanzo un dramma, romanzo confessioni scene, pagine romanzo memorie romanzo	novela novela novela novela episodios novela memorias, novela memorias episodios historias	nuvelă poveste poveste nuvelă nuvelă nuvela roman roman roman roman roman scriere roman


100
75
50
25
0
count
(T1)
(T2)
(T3)
(T4)
Slot
twoContentElement
no
yes

Thanks for your attention!

- Open for questions, comments and contributions!