Text2Story Lusa

A Dataset for Narrative Analysis in European Portuguese News Articles

Sérgio Nunes, Alípio Jorge, Evelin Amorim, Hugo Sousa, António Leal, Purificação Silvano, Inês Cantante, Ricardo Campos

Narratives have been the subject of extensive research across various scientific fields such as linguistics and computer science. However, the scarcity of freely available datasets remains a significant obstacle.

To address this gap, we developed the Text2Story Lusa datasets, which consist of two independent datasets:

Text2Story Lusa and the **Text2Story Lusa Annotated Corpus**.

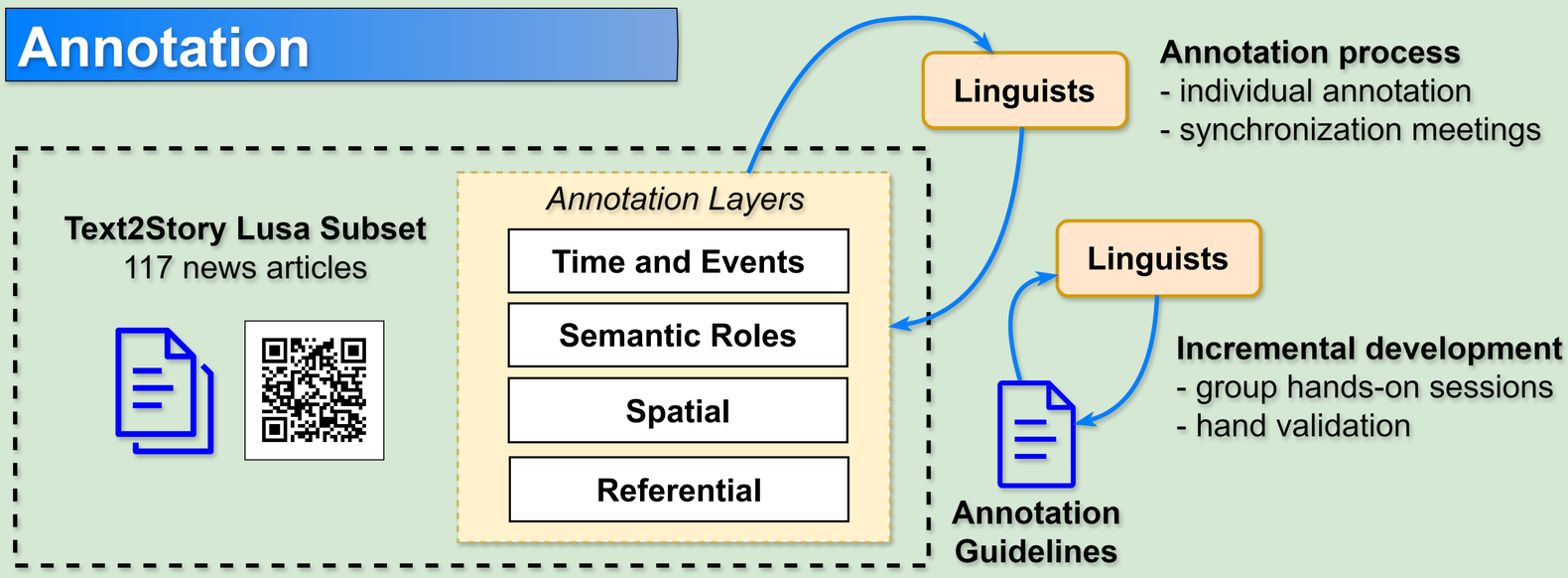
Construction

1st selection criteria:
- size (50 to 200 words)
- language (PT only)
- no repetitions

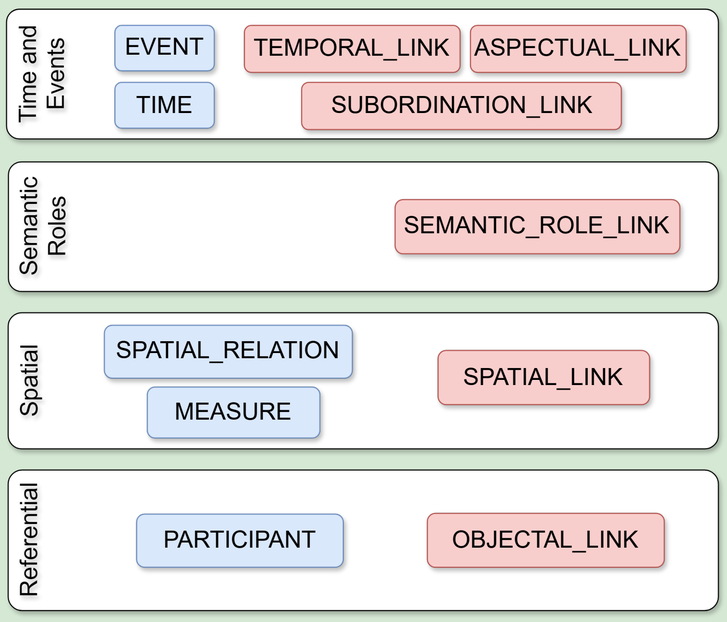
2nd selection criteria:
+ set of keywords

Text2Story Lusa
357 news articles

Annotation



Annotation Scheme



Examples

```
{
  "articles": [
    {
      "id": "102",
      "location": "Lisboa",
      "publication_time": "2020-12-12",
      "headline": "Homem socorrido no Rio Tejo está livre de perigo",
      "content": "Um homem de 68 anos caiu hoje no Rio Tejo, junto ao Cais das
↪  Colunas, […] onde ainda se encontra internado, mas livre de perigo."
    }
  ]
}
```

Listing 1: Text2Story Lusa news article example.

```
T64    Event 930 937 Segundo
A91    Class T64 Reporting
T53    Participant 938 943 a PSP
A286   Lexical_Head T53 Noun
A287   Individuation_Domain T53 Individual
A288   Participant_Type_Domain T53 Org
A289   Involvement T53 1
T54    Participant 945 953 o detido
A290   Lexical_Head T54 Noun
A291   Individuation_Domain T54 Individual
A292   Participant_Type_Domain T54 Per
A293   Involvement T54 1
T21    Time 956 960 hoje
A145   Time_Type T21 Date
A146   TemporalFunction T21 Publication_Time
T22    Event 961 996 presente às Autoridades Judiciárias
A147   Class T22 Occurrence
A148   Event_Type T22 Transition
A149   Pos T22 Verb
A150   Tense T22 Present
A154   Polarity T22 Pos
R86    SRLINK_agent Arg1:T53 Arg2:T64
R67    TLINK_after Arg1:T22 Arg2:T64
R78    SRLINK_patient Arg1:T22 Arg2:T54
R19    TLINK_isIncluded Arg1:T22 Arg2:T21
```

Listing 2: An annotation excerpt, in the BRAT standoff format, from the Text2Story Lusa Dataset. The complete annotated sentence is “Segundo a PSP, o detido é hoje presente às Autoridades Judiciárias.” (Translated: “According to the PSP, the detainee is present today at the Judicial Authorities”).

Characterization

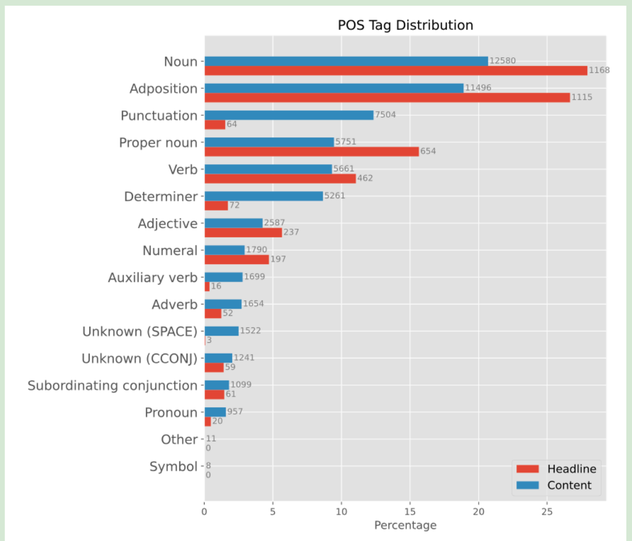


Figure 3: Distribution of POS types in the ‘headline’ and the ‘content’ fields.

Narrative Component	Frequency
Participants	3,530
Events	3,027
Spacial Relations	512
Times	438
Measure	9
Semantic Role Links	4,764
Temporal Links	3,522
Objectal Links	2,224
Qualitative Spatial Link	927
Subordination Link	429
Movement Link	125
Aspectual Link	18

Table 1: Frequency of each narrative component in the Text2Story annotated corpus.

Conclusions

Text2Story Lusa is a linguistic resource comprised of two datasets developed to support research in various fields.

Text2Story Lusa includes a manually curated collection of news articles written in European Portuguese, obtained from the Lusa news agency.

Text2Story Lusa Annotated Corpus includes manual annotations for a subset of the original articles. The annotation adopts a multilayer semantic scheme comprising entity structures and link structures.

INESCTEC · Centro de Linguística da Universidade do Porto · U.PORTO · UNIVERSIDADE BEIRA INTERIOR · fct Fundação para a Ciência e a Tecnologia

LREC-COLING 2024 | Turin, Italy