ELTeC-fra: State of Play

Christof Schöch

ELTeC Zoom Party, Nov. 7, 2020

Universität Trier

Overview

1. How was work on ELTeC-fra organized?
2. Current state of the collection
3. How were texts selected?
4. How did you learn how to use TEI
5. How were texts processed?
6. How were texts processed?
7. What challenges did you encounter?

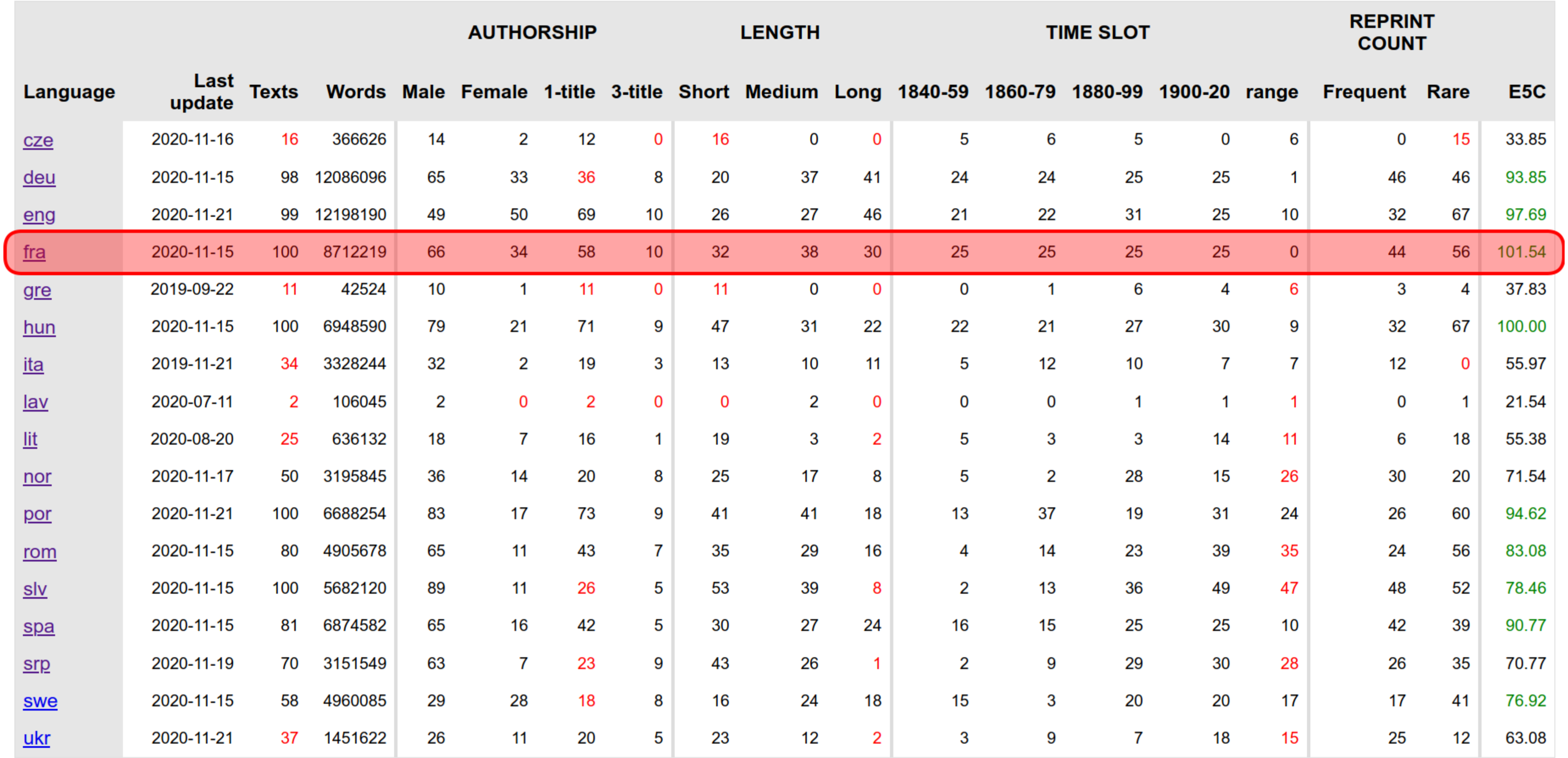
(1) How was work on ELTeC-fra organized?

Team

- Coordination: Christof
- Text selection: Christof
- Encoding strategy: Lou
- Text encoding: Christof, Pia Geißel, Evegnia Fileva
- Python programming: Christof, Johanna Konstanciak, Anne Klee
- Obtaining texts: Christof, Lou, Alexandre Gefen

(2) Current state of the collection

Composition Criteria (E5C)



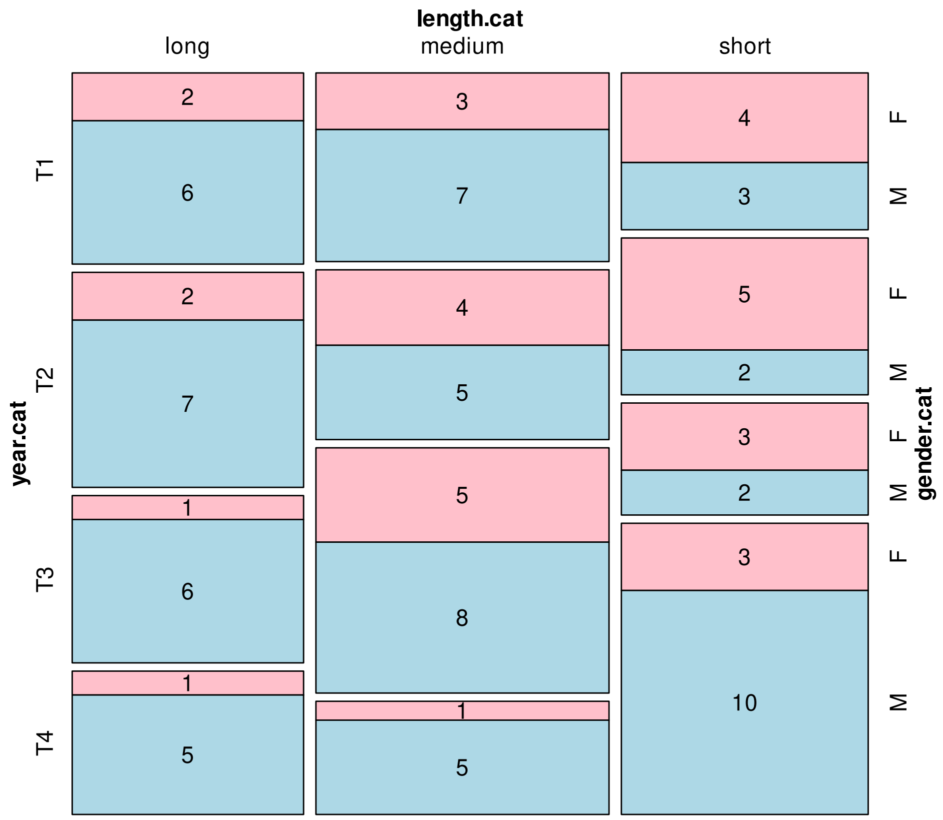
				AUTHORSHIP				LENGTH			TIME SLOT					REPRINT COUNT		
Language	Last update	Texts	Words	Male	Female	1-title	3-title	Short	Medium	Long	1840-59	1860-79	1880-99	1900-20	range	Frequent	Rare	E5C
cze	2020-11-16	16	366626	14	2	12	0	16	0	0	5	6	5	0	6	0	15	33.85
deu	2020-11-15	98	12086096	65	33	36	8	20	37	41	24	24	25	25	1	46	46	93.85
eng	2020-11-21	99	12198190	49	50	69	10	26	27	46	21	22	31	25	10	32	67	97.69
fra	2020-11-15	100	8712219	66	34	58	10	32	38	30	25	25	25	25	0	44	56	101.54
gre	2019-09-22	11	42524	10	1	11	0	11	0	0	0	1	6	4	6	3	4	37.83
hun	2020-11-15	100	6948590	79	21	71	9	47	31	22	22	21	27	30	9	32	67	100.00
ita	2019-11-21	34	3328244	32	2	19	3	13	10	11	5	12	10	7	7	12	0	55.97
lav	2020-07-11	2	106045	2	0	2	0	0	2	0	0	0	1	1	1	0	1	21.54
lit	2020-08-20	25	636132	18	7	16	1	19	3	2	5	3	3	14	11	6	18	55.38
nor	2020-11-17	50	3195845	36	14	20	8	25	17	8	5	2	28	15	26	30	20	71.54
por	2020-11-21	100	6688254	83	17	73	9	41	41	18	13	37	19	31	24	26	60	94.62
rom	2020-11-15	80	4905678	65	11	43	7	35	29	16	4	14	23	39	35	24	56	83.08
slv	2020-11-15	100	5682120	89	11	26	5	53	39	8	2	13	36	49	47	48	52	78.46
spa	2020-11-15	81	6874582	65	16	42	5	30	27	24	16	15	25	25	10	42	39	90.77
srp	2020-11-19	70	3151549	63	7	23	9	43	26	1	2	9	29	30	28	26	35	70.77
swe	2020-11-15	58	4960085	29	28	18	8	16	24	18	15	3	20	20	17	17	41	76.92
ukr	2020-11-21	37	1451622	26	11	20	5	23	12	2	3	9	7	18	15	25	12	63.08

Summary produced: 2020-11-24

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

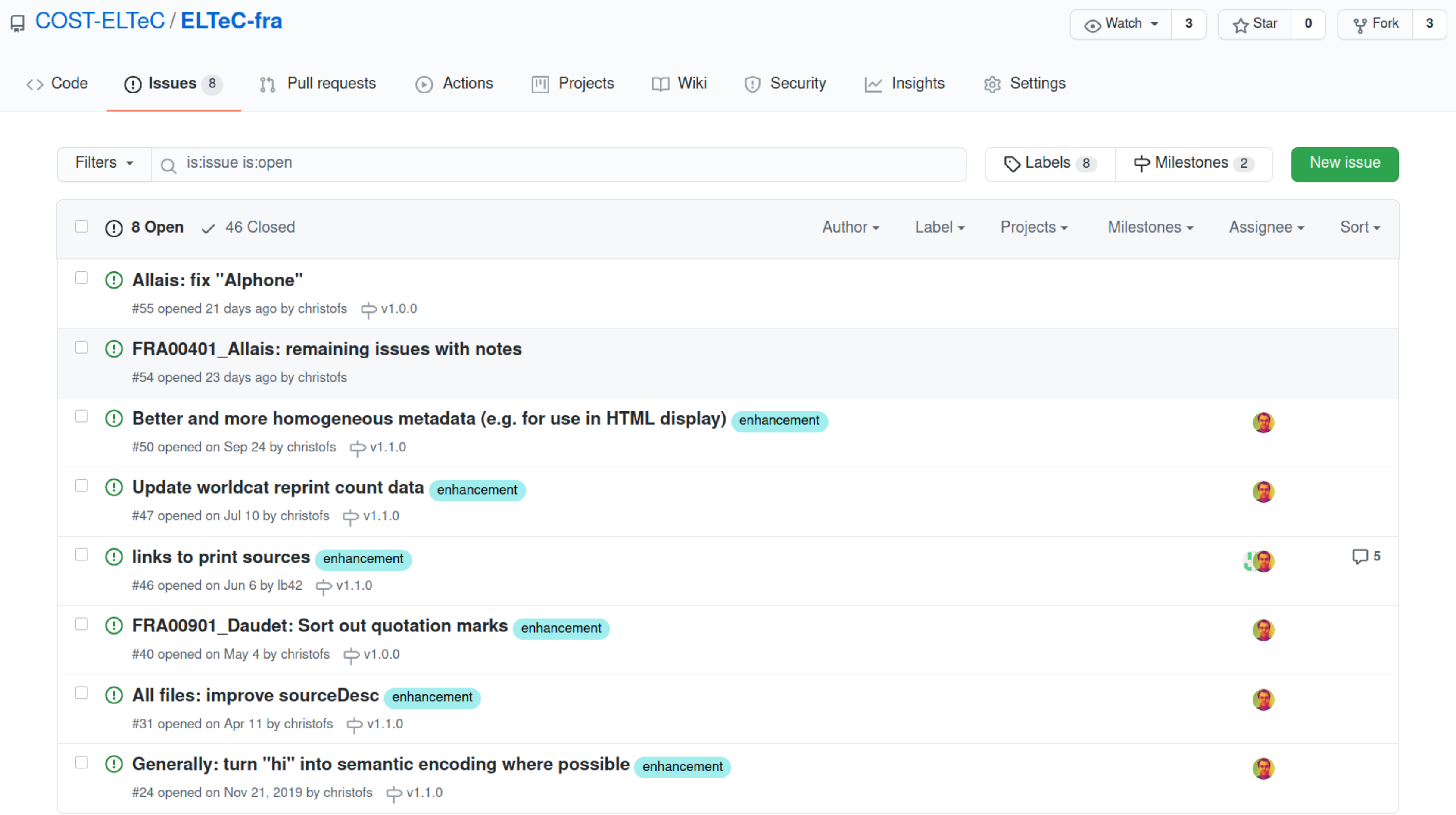
Composition Criteria (mosaic plot)

Title counts for each balance criterion



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

Open Issues



Allais: fix "Alphone"
#55 opened 21 days ago by christofs v1.0.0

FRA00401_Allais: remaining issues with notes
#54 opened 23 days ago by christofs

Better and more homogeneous metadata (e.g. for use in HTML display) enhancement
#50 opened on Sep 24 by christofs v1.1.0

Update worldcat reprint count data enhancement
#47 opened on Jul 10 by christofs v1.1.0

links to print sources enhancement 5
#46 opened on Jun 6 by lb42 v1.1.0

FRA00901_Daudet: Sort out quotation marks enhancement
#40 opened on May 4 by christofs v1.0.0

All files: improve sourceDesc enhancement
#31 opened on Apr 11 by christofs v1.1.0

Generally: turn "hi" into semantic encoding where possible enhancement
#24 opened on Nov 21, 2019 by christofs v1.1.0

https://github.com/COST-ELTeC/ELTeC-fra/issues

(3) How were texts selected?

Selection

Selection

- First phase: quantity
  - low-hanging fruit = texts existing in marked-up form
  - main sources: ELG, CLiGS, other EPUB sources

Selection

- First phase: quantity
    - low-hanging fruit = texts existing in marked-up form
    - main sources: ELG, CLiGS, other EPUB sources
- Second phase: quality
    - strategic additions: optimize E5C score
    - main sources: BnF, OBVIL

(4) How did you learn how to use TEI

Various training strategies

- Pia Geißel participated in the Budapest Training School
- Evgenia Fileva was studying a CL/DH programme in Trier
- Christof learned from Lou's feedback

(5) How were texts processed?

”Manual” pipeline

1. ~~Scanning and OCR~~

”Manual” pipeline

1. ~~Scanning and OCR~~
2. EPUB from ELG or BnF (XML or TXT: skip to #4)

”Manual” pipeline

1. ~~Scanning and OCR~~
2. EPUB from ELG or BnF (XML or TXT: skip to #4)
3. Calibre: read EPUB, transform to Markdown

”Manual” pipeline

1. ~~Scanning and OCR~~
2. EPUB from ELG or BnF (XML or TXT: skip to #4)
3. Calibre: read EPUB, transform to Markdown
4. Python scripts: transform to ELTeC XML-TEI, mostly using RegEx

”Manual” pipeline

1. ~~Scanning and OCR~~
2. EPUB from ELG or BnF (XML or TXT: skip to #4)
3. Calibre: read EPUB, transform to Markdown
4. Python scripts: transform to ELTeC XML-TEI, mostly using RegEx
5. Atom: fix encoding errors, add metadata, validate

”Manual” pipeline

1. ~~Scanning and OCR~~
2. EPUB from ELG or BnF (XML or TXT: skip to #4)
3. Calibre: read EPUB, transform to Markdown
4. Python scripts: transform to ELTeC XML-TEI, mostly using RegEx
5. Atom: fix encoding errors, add metadata, validate
6. Atom: spell-check

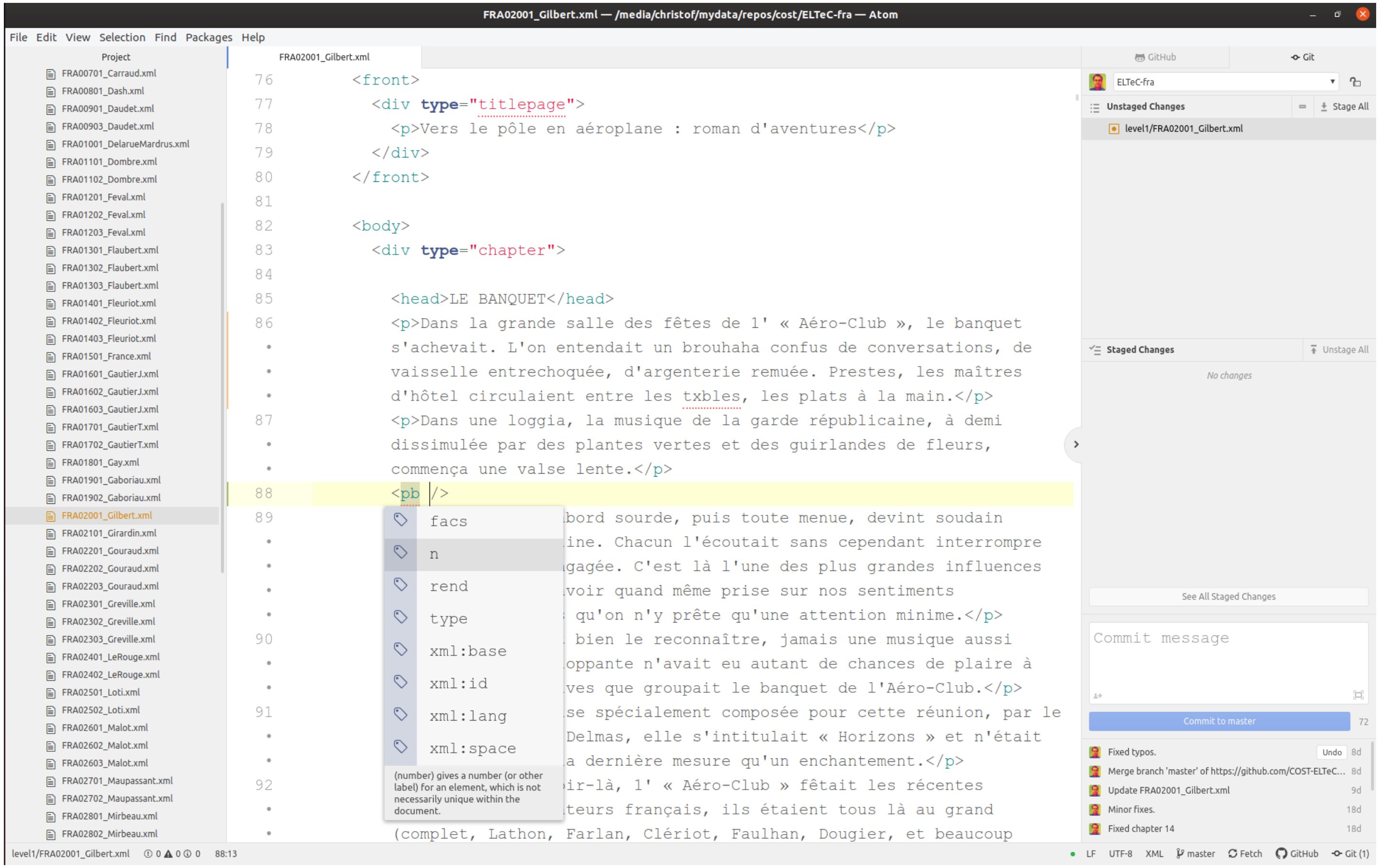
”Manual” pipeline

1. ~~Scanning and OCR~~
2. EPUB from ELG or BnF (XML or TXT: skip to #4)
3. Calibre: read EPUB, transform to Markdown
4. Python scripts: transform to ELTeC XML-TEI, mostly using RegEx
5. Atom: fix encoding errors, add metadata, validate
6. Atom: spell-check
7. Atom: upload to repository

”Manual” pipeline

1. ~~Scanning and OCR~~
2. EPUB from ELG or BnF (XML or TXT: skip to #4)
3. Calibre: read EPUB, transform to Markdown
4. Python scripts: transform to ELTeC XML-TEI, mostly using RegEx
5. Atom: fix encoding errors, add metadata, validate
6. Atom: spell-check
7. Atom: upload to repository
8. TXM: linguistic annotation

Atom: game changer



```
      <front>
        <div type="titlepage">
          <p>Vers le pôle en aéroplane : roman d'aventures</p>
        </div>
      </front>

      <body>
        <div type="chapter">

          <head>LE BANQUET</head>
          <p>Dans la grande salle des fêtes de l' « Aéro-Club », le banquet
          s'achevait. L'on entendait un brouhaha confus de conversations, de
          vaisselle entrechoquée, d'argenterie remuée. Prestes, les maîtres
          d'hôtel circulaient entre les txbles, les plats à la main.</p>
          <p>Dans une loggia, la musique de la garde républicaine, à demi
          dissimulée par des plantes vertes et des guirlandes de fleurs,
          commença une valse lente.</p>
          <pb />
                              bord sourde, puis toute menue, devint soudain
                              ine. Chacun l'écoutait sans cependant interrompre
                              gagée. C'est là l'une des plus grandes influences
                              voir quand même prise sur nos sentiments
                              qu'on n'y prête qu'une attention minime.</p>
                              bien le reconnaître, jamais une musique aussi
                              oppante n'avait eu autant de chances de plaire à
                              ves que groupait le banquet de l'Aéro-Club.</p>
                              se spécialement composée pour cette réunion, par le
                              Delmas, elle s'intitulait « Horizons » et n'était
                              a dernière mesure qu'un enchantement.</p>
                              ir-là, l' « Aéro-Club » fêtait les récentes
                              teurs français, ils étaient tous là au grand
          (complet, Lathon, Farlan, Clériot, Faulhan, Dougier, et beaucoup
```

facs
n
rend
type
xml:base
xml:id
xml:lang
xml:space

(number) gives a number (or other label) for an element, which is not necessarily unique within the document.

Fixed typos.	8d
Merge branch 'master' of https://github.com/COST-ELTeC...	8d
Update FRA02001_Gilbert.xml	9d
Minor fixes.	18d
Fixed chapter 14	18d

- Validation, spell-check, push/pull
- Details: https://dragonfly.hypotheses.org/1127

(6) What challenges did you encounter?

Challenges

Challenges

1. Composition: For some metadata combinations, there are just very few texts (French: T4 + female + long); production or availability?

Challenges

1. Composition: For some metadata combinations, there are just very few texts (French: T4 + female + long); production or availability?
2. Copyright issues: one novel from 1884 had to be excluded after being selected (author died in 1953)

Challenges

1. Composition: For some metadata combinations, there are just very few texts (French: T4 + female + long); production or availability?
2. Copyright issues: one novel from 1884 had to be excluded after being selected (author died in 1953)
3. Text quality: some novels were produced from faulty scans with missing characters at the margins (solution: replace)

Challenges

1. Composition: For some metadata combinations, there are just very few texts (French: T4 + female + long); production or availability?
2. Copyright issues: one novel from 1884 had to be excluded after being selected (author died in 1953)
3. Text quality: some novels were produced from faulty scans with missing characters at the margins (solution: replace)
4. Reprint count: The 'worldcat.py' script just never produced really reliable data

Danke!