A CORPUS OF NARRATIVES RELATED TO LUXEMBOURG FOR THE PERIOD 1945-1975

TIR 2017

Olivier Parisot, Thomas Tamisier

LUXEMBOURG
INSTITUTE
OF SCIENCE
AND TECHNOLOGY
LIST

SUMMARY

- Context
- Data sources
- Corpus construction
- Corpus enrichment
- Corpus analysis
- Conclusion

SUMMARY

- **Context**
- Data sources
- Corpus construction
- Corpus enrichment
- Corpus analysis
- Conclusion

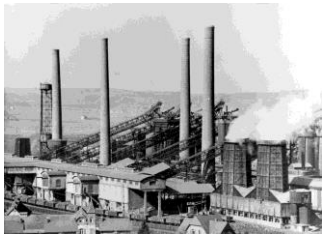
CONTEXT (1/2)

The LOCAlised LEgacy project (LOCALE)

- A project funded by the FNR
  - *https://www.fnr.lu/projects/localised-legacies-2/*

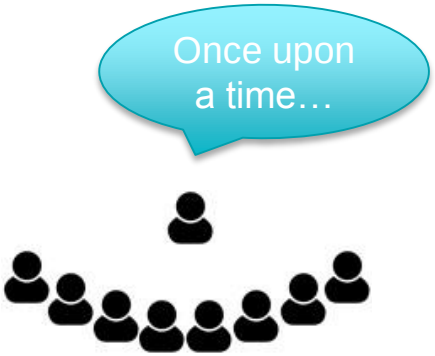
Luxembourg National Research Fund

- Goal: gather stories and narratives occurred after the WWII, between the years of **1945-1975**, and in **Luxembourg and its region**.

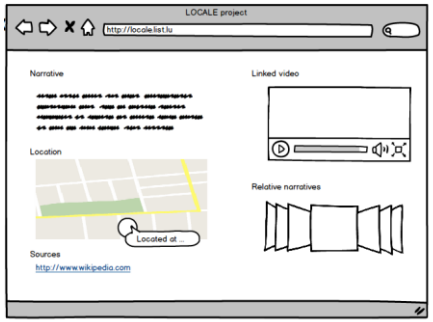


CONTEXT (2/2)

The LOCAlised LEgacy project (LOCALE)



- Workshops with elderly people were conducted
  (by using pictures, books, maps).
  -> Initial effort is required to trigger **the recall of stories**.



- Partners

SUMMARY

- Context
- **Data sources**
- Corpus construction
- Corpus enrichment
- Corpus analysis
- Conclusion

DATA SOURCES (1/2)

Getting stories from public institutions?

- Centre National de l'Audiovisuel
  - Provides multimedia material for the considered period, but the textual data are essentially meta-data that were added a posteriori to describe videos or audios.

- Centre virtuel de la connaissance sur l'Europe:
  - Provides official historical documents about European construction (legal texts, treaties, European directives, etc.).

- Related works from the Luxembourg university
  - *P. Gilles and E. Ziegler (2013)*
    - *"The historical luxembourgish bilingual public notices database"*

DATA SOURCES (2/2)

Getting stories from home retirements?

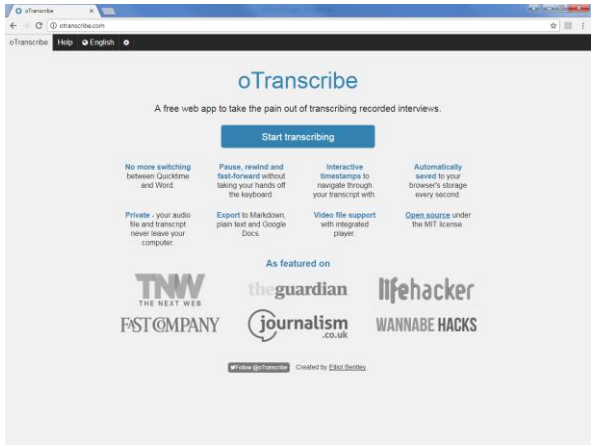
Google Cloud Platform

KonviktsGaart
Résidence Grande-Duchesse Joséphine-Charlotte

hello

Once upon a time…

Speechnotes



SUMMARY

- Context
- Data sources
- **Corpus construction**
- Corpus enrichment
- Corpus analysis
- Conclusion

CORPUS CONSTRUCTION (1/4)

Our approach

- Getting *fr/lu/de/en* narratives about Luxembourg from
  - Wikipedia
    - To benefit both of the multilingual capabilities and the work of the large community of contributors.
      - https://lb.wikipedia.org/wiki/Esch-Uelzecht
      - https://fr.wikipedia.org/wiki/Esch-sur-Alzette
      - …
  - Wikidata & Open Data Portal (https://data.public.lu)
    - To build a seed words list composed of Luxembourgish
      - well-known locations (i.e. *Kirchberg*),
      - persons (i.e. *Viviane Reding*)
      - companies (i.e. *CargoLux*)
      - etc.
- Development of a JAVA 8 software
  - *https://git.list.lu/eScience/prototext (currently: restricted access ☹)*

CORPUS CONSTRUCTION (2/4)

Algorithm

Algorithm 1 Pseudo-code to retrieve a multilingual corpus of narratives from Wikipedia and Wikidata.

```
narratives ← empty list
for lang in (lb,fr,de) do
    entities ← retrieve the entities names related to Luxembourg from Wikidata (according to lang)
    for all entities do
        url ← build the URL of the Wikipedia query (according to lang)
        content ← retrieve the content from url
        sentences ← split content into sentences
        for sentence in sentences do
            if sentence is syntactically and grammatically correct then
                location ← extract location from sentence
                if location ≠ null and location in Luxembourg then
                    date ← extract date from sentence
                    if date ≠ null and 1945 ≤ date ≤ 1975 then
                        (latitude, longitude) ← get coordinates for location
                        add (sentence, lang, date, location, latitude, longitude) into narratives
                    end if
                end if
            end if
        end for
    end for
end for
return narratives
```

CORPUS CONSTRUCTION (3/4)

Technical aspects

- Wikidata SPARQL interface (https://query.wikidata.org/)

```
SELECT ?a ?aLabel ?birth_date WHERE
{
 ?a wdt:P27 wd:Q32 . # Q32 -> Luxembourg
 ?a p:P569/psv:P569 ?birth_date_node .
 ?birth_date_node wikibase:timeValue ?birth_date .
 FILTER (year(?birth_date) >= 1945) .
 FILTER (year(?birth_date) <= 1975)
 SERVICE wikibase:label {bd:serviceParam  wikibase:language "lb" .}
}
```

- Wikipedia REST API

```
https://[LANG].wikipedia.org/w/api.php?action=opensearch&search=[QUERY]
                                        &limit=[COUNT]&namespace=0&format=json
```

- LanguageTool (style and grammar checker for 25+ languages)

```
https://github.com/languagetool-org/languagetool
```

CORPUS CONSTRUCTION (4/4)

Data model

- Gathered stories are stored in a JSON file:

```
{
    "date": "1975",
    "locationName": "Echternach",
    "text": "Since 1975, Echternach has been the site of an International Music Festival, held annually in May and June.",
    "source": "wikipedia",
    "lang": "en",
    …
}
```

- As a result:
  - A multilingual corpus of 1104 narratives related to the Luxembourg during the 1945-1975 period.

TABLE IV
QUANTITATIVE DESCRIPTION OF THE GENERATED CORPUS. ENGLISH STORIES COUNT IS PRESENTED FOR INFORMATION PURPOSE ONLY.

Language	Count of stories	Stories characters count
French	266	min: 38 max: 585 avg: 139
German	240	min: 35 max: 3585 avg: 404
Luxembourgish	598	min: 32 max: 1360 avg: 101
English	*350*	*min: 23 max: 3092 avg: 286*

SUMMARY

- Context
- Data sources
- Corpus construction
- **Corpus enrichment**
- Corpus analysis
- Conclusion

CORPUS ENRICHMENT (1/3)

Geolocalization

- Getting longitude/latitude


```
{
    "date": "1975",
    "locationName": "Echternach",
    "text": "Since 1975, Echternach has been the site of an International Music Festival,
    held annually in May and June.",
    "source": "wikipedia",
    "lang": "en",
    "longitude": 6.4175635,
    "latitude": 49.8114133
    …
}
```


- Google Maps Geocoding API

https://maps.googleapis.com/maps/api/geocode/json?key=[KEY]&address=[ADDRESS]

CORPUS ENRICHMENT (2/3)

Text mining

- Named Entity Recognition + query to knowledge base:
  - *From 1970, Alexander Mullenbach taught piano at the Conservatoire de Luxembourg.*

- Natural Language Understanding API (formerly AlchemyAPI)
  - https://www.ibm.com/watson/services/natural-language-understanding/

- Google Knowledge Graph

CORPUS ENRICHMENT (3/3)

Narratives classification

- Classification:
  - Classes: inspired by the IAB taxonomy
  - For each language: lu/fr/de/en

Percentage	Category
21.9%	business and industrial
18.5%	travel
10.7%	law, govt and politic
9.5%	art and entertainment
6.7%	home and garden
5.0%	society
5.0%	science

From 1970, Alexander Mullenbach taught piano at the Conservatoire de Luxembourg.

-> leads to the /art and entertainment/music category.

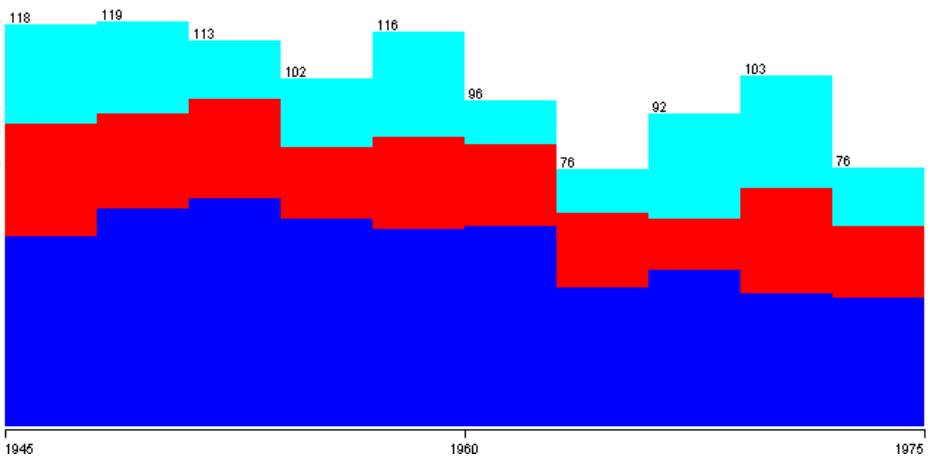
- DISCO
  - *Peter Kolb (2009):*
    - *Experiments on the difference between semantic similarity and relatedness, 17th Nordic Conference on Computational Linguistics - NODALIDA '09.*

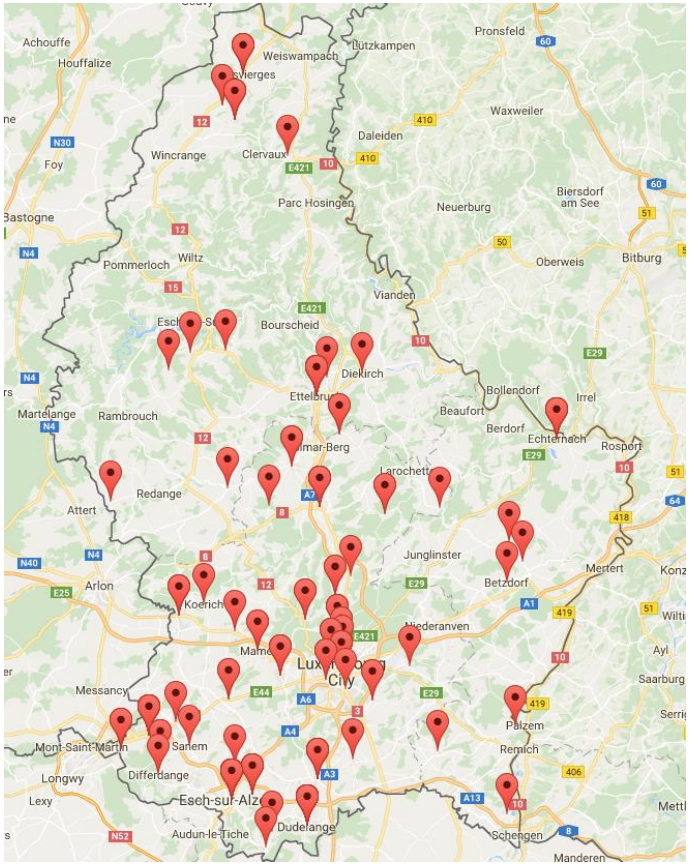
SUMMARY

- Context
- Data sources
- Corpus construction
- Corpus enrichment
- **Corpus analysis**
- Conclusion

CORPUS ANALYSIS (1/3)

Period coverage & Geographical coverage

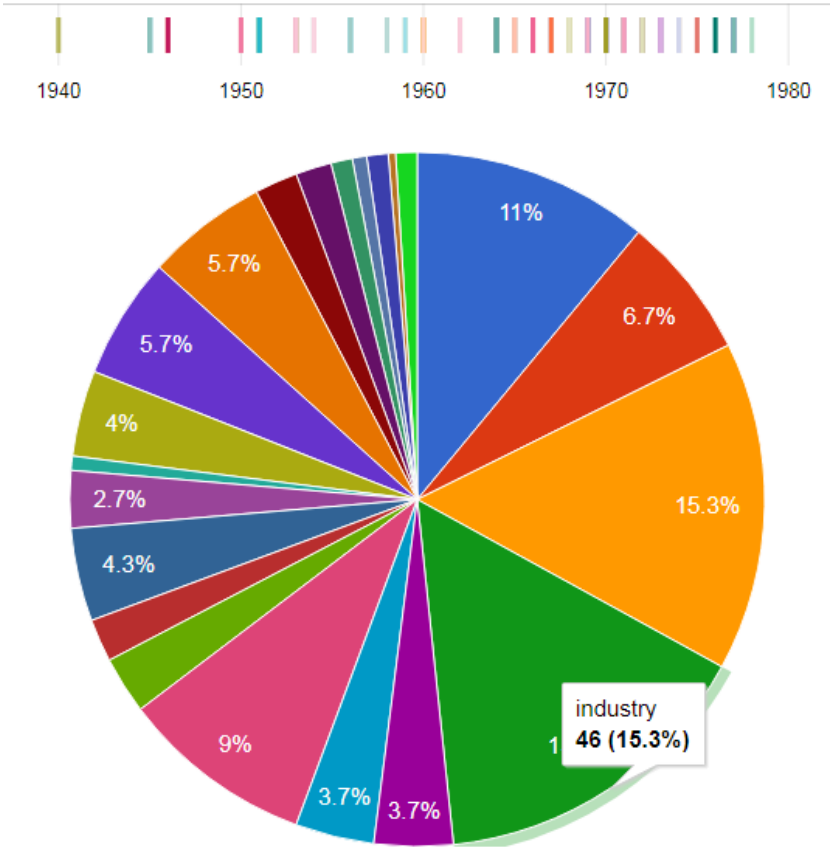




CORPUS ANALYSIS (2/3)

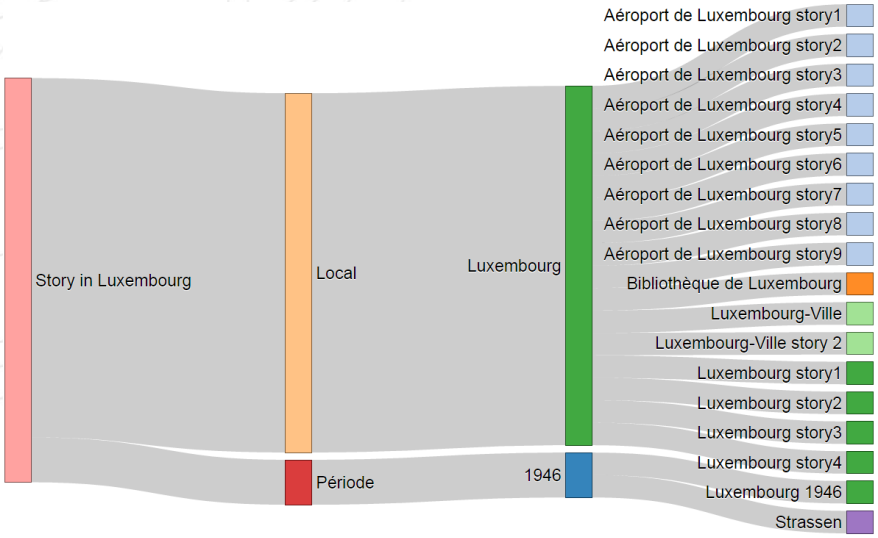
Most frequent words & Categories

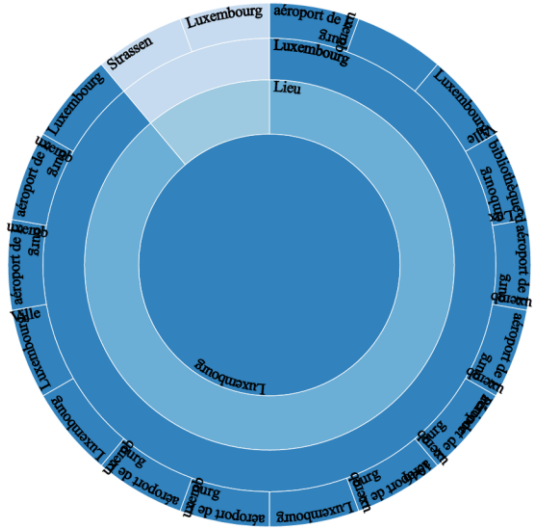
aeroport airlines Arbed aviation Capitole centre CFL chateau chemins cinema Cite commence compagnie construction construit cours creee derniere emetteur entreprise fer fermeture fondation fondee guerre industrie jean ligne lors monde mondiale nationale officielle ouverture parc partir pont premiere prince production projet reseau service societe synagogue trafic travaux usine villa visite

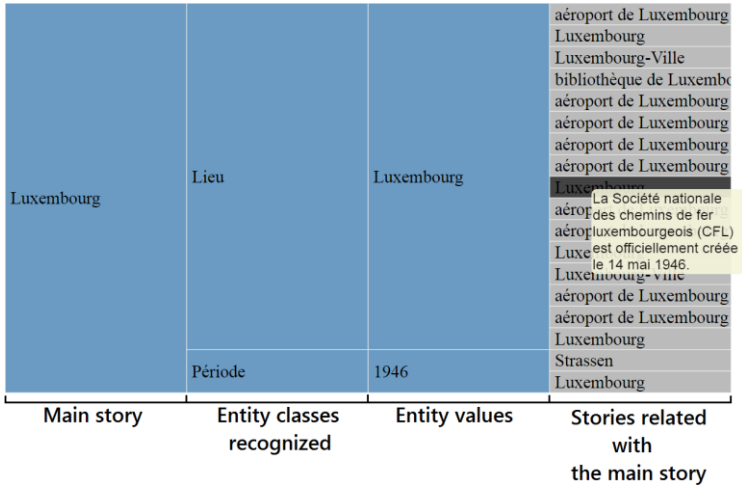


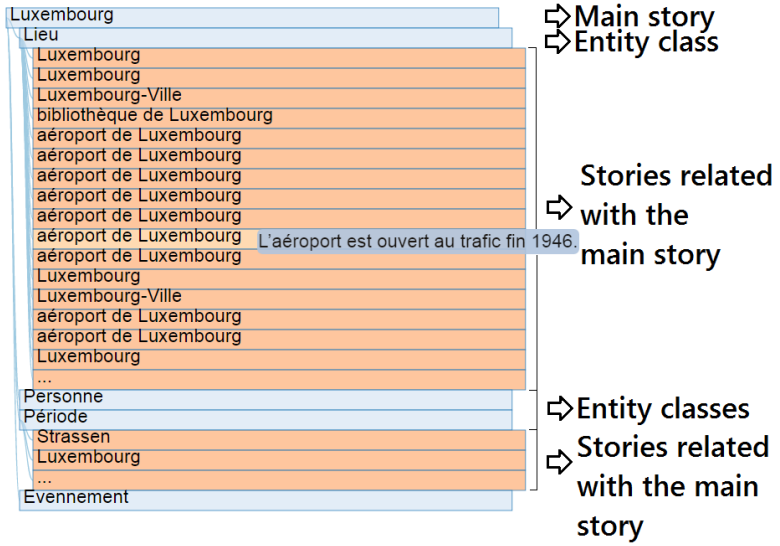
CORPUS ANALYSIS (3/3)

Hierarchical dataviz with D3JS









SUMMARY

- Context
- Data sources
- Corpus construction
- Corpus enrichment
- Corpus analysis
- **Conclusion**

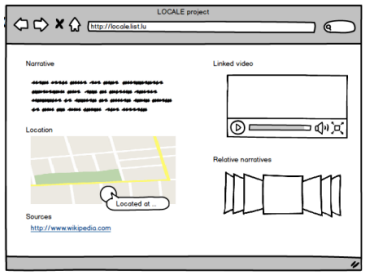
CONCLUSION

Done / work in progress

- Goal: facilitate the recall of stories about the Luxembourg (1945/1975).

- Done:
  - Construction of a multilingual corpus of stories from the web.

- Work in progress:
  - A dedicated back-end to manage narratives and the associated meta-data
    - *Pierrick Bruneau, Olivier Parisot, Thomas Tamisier (2017)*
      - *Storing and Processing Personal Narratives in the context of Cultural Legacy Preservation*
  - Dataviz to analyse stories
    - *Paulo Carvalho, Olivier Parisot, Thomas Tamisier (2017)*
      - *Using Visualisation Techniques to Acquire a Better Understanding of Storytelling for Cultural Heritage*
  - Web & mobile clients apps



CONCLUSION

Perspectives

- Applying Natural Language Generation (NLG) on the corpus to generate questions
  - Example:
    - *Do you have a story related to the marriage of Grand-Duc Jean with Princess Jos´ephine-Charlotte in 1953?*
  - More Than One Story card game
    - http://www.simrishamn.se/mtos

THANK YOU!

QUESTIONS?

BONUS

(Semantic) Similarity of narratives

- Multi-dimensional scaling projection + colors