A corpus of narratives related to Luxembourg for the period 1945-1975

Olivier Parisot, Thomas Tamisier
Luxembourg Institute of Science and Technology
41, rue du Brill L-4422 Luxembourg
email: olivier.parisot@list.lu

***Abstract*—Acquiring stories and narratives about past periods is a challenge for cultural heritage preservation. In this context, we present a method to obtain from the web a corpus of texts related to the period of 1945-1975 in Luxembourg. Extracted texts are accompanied by meta-data that facilitate their integration by tier applications. As a use-case, this corpus will be used in a software that aims at helping elderly people to recall and share anecdotal stories about this period.**

I. Introduction

The Grand Duchy of Luxembourg, is a small country in Western Europe (the population was estimated to 313970 in 1960 and 576249 in 2016). Due to its strategical geographic location (at the border of Belgium, Germany and France), the country have been impacted by World War II and has a central role in the European Union.

In the context of a research project, we are developing a software platform to store, share and reuse personal narratives of citizens about the time of the European construction in Luxembourg [1].

Preliminary workshops with elderly people (between 70 and 85 years old) were conducted. An important set of narratives under textual and oral formats were collected. Nevertheless, we have observed during the workshops that initial effort is required to trigger the recall of stories by elderly people (for example: by providing them contextual information about their personal anecdotes). To this end, we need to pre-fill the database of our software by collecting narratives about Luxembourg history related to:

- Economical development of the country: industry (steel production until 1970s), banking (important financial sector from 1960s), telecommunications (*Compagnie luxembourgeoise de télédiffusion – Radio Television Luxembourg*), etc.
- Transport infrastructures development: civil aviation (Luxembourg-Findel airport extension), etc.
- European integration: installation of official institutions such as the European Court of Justice at Luxembourg city in 1952, etc.
- art and culture, sports, etc.

In this paper, we present a method to obtain from the web a corpus of textual narratives related to the period of 1945-1975 in Luxembourg.

The rest of this article is organized as follows. Firstly, related work about narratives corpus construction is discussed in Section II. Then, we define a data model to store narratives in Section III, and we propose a method to extract such stories from Wikipedia in Section IV. Next, we briefly present a quantitative and qualitative analysis of the obtained corpus (Section V). Finally, we conclude with the presentation of several perspectives (Section VI).

II. Related works

To the best of our knowledge, the desired narratives corpus does not exist, and there is no work that directly addresses the problem of building a corpus of narratives for Luxembourg for a given period. Some public Luxembourgish organisations provide data about Luxembourg history:

- The *Centre National de l'Audiovisuel* [1] institution 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.
- The *Centre virtuel de la connaissance sur l'Europe* (recently attached to University of Luxembourg and renamed as 'CVCE.eu by UNI.lu') [2] provides official historical documents about European construction (legal texts, treaties, European directives, etc.).

From the academical point of view, [2] describes a corpus containing French-German Luxembourgish public notices covering the years from 1795 to 1920. In this case, the type of data are interesting, but the covered period is not suitable for our work.

After analyzing related work in corpus construction, we identified two major steps.

In a first phase, content should be found and extracted from the World Wide Web by applying web crawling or by invoking web search engines. Even if data are potentially accessible to everyone, getting valuable textual data from the web is not straightforward [3]. For example, various actions (like cleaning HTML, Javascript, headers, disclaimers, etc.) are mandatory to obtain exploitable content [4].

The second phase consists in analyzing and annotating the extracted data by applying methods to extract temporal [5] and geographical information [6], [7], [8] from raw text. Stories matching the considered period and location can then be retained.

[1] http://www.cna.public.lu
[2] http://www.cvce.eu

For this work, we have applied this approach to build a corpus from the web, in such a way that narratives are related to Luxembourg between 1945 and 1975.

III. Corpus model

Various works have proposed models to store factual (non-fictive) narratives. In [9], the authors aim at annotating and linking stories to facilitate their further analyze (for instance, highlighting the connection of different stories).

For this work, we have defined a simple flat model to describe narratives by these fields:

- The raw text (example: *"After acquiring the Villa Vauban in 1949, the City of Luxembourg adapted it for rental by the ECSC Court Of Justice."*). Ideally, the text should be composed by one or several correct sentences (syntactically and grammatically).
- Luxembourg has three different official languages: Luxembourgish, French, German). Due to the multilingual culture of Luxembourg, we have to consider that stories can be written in those languages.
- The date, as accurate as possible (at least the year).
- The location, as accurate as possible (at least the city).
- The source (i.e. the URL from which the story was obtained).

In practice, a story can refer to several locations and dates: models like [9] manage this aspect (when the narrative tells a journey, for example). Nevertheless, we avoid taking this aspect into account to build the corpus, and we prefer to get stories for which it is possible to infer a date and a location.

IV. Corpus generation

Basically, our method consists in retrieving textual data from specific data sources and selecting correct sentences that exactly contain dates and locations (Algorithm 1).

In a recent work, the authors propose to build a corpus of German sentences from the web by parsing a wide range of German and Austrian websites [4]. In our case, we have preferred to only select two data sources:

- Wikipedia: to benefit both of the multilingual capabilities and the work of the large community of contributors. As an example, we can make an overlap between the contributions independently of the language (ex: *https://lb.wikipedia.org/wiki/Esch-Uelzecht* vs *https://fr.wikipedia.org/wiki/Esch-sur-Alzette*).
- Wikidata: an open knowledge base that provides a structured version of Wikipedia's data. Even if recent works have shown that Wikidata is still young and immature [10], this platform provides a convenient query mechanism based on the standard SPARQL language.

More precisely, the approach is implemented in a JAVA standalone tool and is composed by several steps.

First of all, we produce a list of entities that are related to Luxembourg (Alg. 1, line 3). This list (sometimes called *seed words list* [11]) is composed of Luxembourgish well-known locations (i.e. *Kirchberg*) persons (i.e. *Viviane Reding*) and companies (i.e. *Spuerkeess*). To obtain it, we aggregate data from different sources such as the Wikidata query service [12] and the Luxembourgish Open Data platform [3]. As an example, we have executed a SPARQL query to obtain the list of people born in Luxembourg during the considered period (Table I).

TABLE I
A SPARQL query to retrieve from Wikidata all people born at Luxembourg between 1945 and 1975. In April 2017, the query returned a list of 677 persons.

```
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" .
 }
}
```

TABLE II
The URL pattern used to retrieve *[COUNT]* results from Wikipedia for a given query (*[QUERY]*) in a given language (*[LANG]*).

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

Next, we invoke the search engine of the Wikipedia platform (Alg. 1, line 5) by using the appropriate syntax (Table II). As a result, we obtain a list of URLs matching with the query. For example, if the query is *Arbed* (the name of an old Luxembourgish company in charge of producing steel) for the Luxembourgish language, then we obtain the following URLs:

- *https://lb.wikipedia.org/wiki/Arbedsgebai*
- *https://lb.wikipedia.org/wiki/Centrale_thermique_Arbed_Terres-rouges*
- etc.

Then, we retrieve the pages contents and split the text into sentences (this approach was applied in recent works such as [4]). During the parsing, some pieces of text were syntactically and/or grammatically incorrect. To detect and filter them (Alg. 1, line 9), we have used the style and grammar checker provided by the LanguageTool, a JAVA open source proofreading library that is widely used (i.e. OpenOffice, LibreOffice, etc.) [13]. This step allows to keep only the *parsable* sentences (as suggested by [4]).

[3] https://data.public.lu

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

1: $narratives \leftarrow$ empty list
2: **for** $lang$ in (lb,fr,de) **do**
3: $entities \leftarrow$ retrieve the entities names related to Luxembourg from Wikidata (according to $lang$)
4: **for all** $entities$ **do**
5: $url \leftarrow$ build the URL of the Wikipedia query (according to $lang$)
6: $content \leftarrow$ retrieve the content from url
7: $sentences \leftarrow$ split $content$ into sentences
8: **for** $sentence$ in $sentences$ **do**
9: **if** $sentence$ is syntactically **and** grammatically correct **then**
10: $location \leftarrow$ extract $location$ from $sentence$
11: **if** $location \neq null$ **and** $location$ in Luxembourg **then**
12: $date \leftarrow$ extract $date$ from $sentence$
13: **if** $date \neq null$ **and** $1945 \leq date \leq 1975$ **then**
14: $(latitude, longitude) \leftarrow$ get coordinates for $location$
15: add $(sentence, lang, date, location, latitude, longitude)$ into $narratives$
16: **end if**
17: **end if**
18: **end if**
19: **end for**
20: **end for**
21: **end for**
22: **return** $narratives$

After that, we select sentences containing date and location that match the desired criteria.

For the location detection (Alg. 1, line 10), we prepare list of $\tilde{1}745$ well-known Luxembourgish locations (by using the Wikidata service and by taking into account that cities and places are written differently according to the language). In order to avoid confusion between persons and location (for example: *Roger Bour* VS *the Bour Luxembourgish city*), we apply a Named Entity Recognition phase to identify terms that are considered as persons [14]. This step invokes the Natural Language Understanding API (formerly AlchemyAPI), a *Deep Learning* multilingual online tool to apply advanced text mining [15].

For the date extraction (Alg. 1, line 12), we use heuristic aiming at extracting the most precise date (at least the year, day/month/year if possible). Fuzzy dates expressions such as '*in the 1970s*' are considered as well during the extraction. As for the location detection, this phase takes into account that the date can be written in different formats and languages.

Finally, we obtain the latitude and the longitude by using the Google Maps Geocoding API [4] (Alg. 1, line 14). This service provides – through a REST interface – the geographical coordinates from free-text, e.g. *Athénée de Luxembourg*.

In practice, we have ran the implementation of the algorithm (based on Wikipedia's and Wikidata's data of April 2017). At the end, the obtained corpus is a JSON file containing stories for which each required field is defined (Table III).

[4]https://developers.google.com/maps/documentation/geocoding/

TABLE III
AN ENGLISH STORY AND THE ASSOCIATED META-DATA IN THE JSON FORMAT.


```
{
"date": "1962",
"locationName": "Luxembourg's Nouvel
    Athenee",
"text": "Michel Stoffel completed two
    mosaics in Luxembourg's Nouvel Athenee
     in 1962 and became a member of the
    arts and literature section of Grand
    Ducal Institute.",
"source": "https://en.wikipedia.org/wiki/
    Luxembourg_art",
"latitude": 49.60423979999999,
"longitude": 6.1107796,
"language": "en"
}
```


V. CORPUS ANALYSIS

From a quantitative point of view, the generated corpus contains 1104 stories that are written in Luxembourgish, French and German (Table IV). According to the results, the corpus is mostly composed of Luxembourgish stories.

Without being a surprise, most of the stories are short because they are essentially composed by one sentence ($\approx$ 100 characters) (Table IV). As an example, Table III shows a story related to the *Athénée de Luxembourg*, an academic institution founded in 1817 where well-known Luxembourgish politicians were scholarized.

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*

Fig. 1. A Google Map showing the localization of the narratives that are written in Luxembourgish. A point can represent several stories if they are placed at the same location.

We obtained stories covering the whole country. To illustrate this, Figure 1 shows the geographical repartition of the Luxembourgish stories. In this picture, we can see that the capital *Luxembourg City* and the center/south west of the country concentrate the stories.

Regarding the time dimension, the corpus covers the period from 1945 to 1975 in a balanced way (Figure 2).

Moreover, the stories are related to various topics. As an example, we can observe the most frequent terms with a simple tag cloud (Figure 3). At first glance, most frequent terms mostly refer to *industry* and *transport*.

A deeper analysis is possible by classifying each story. To this end, we have applied again the Natural Language Understanding API [15]. For example, the analysis of the following short story (coming from Wikipedia)

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

leads to the */art and entertainment/music* category.

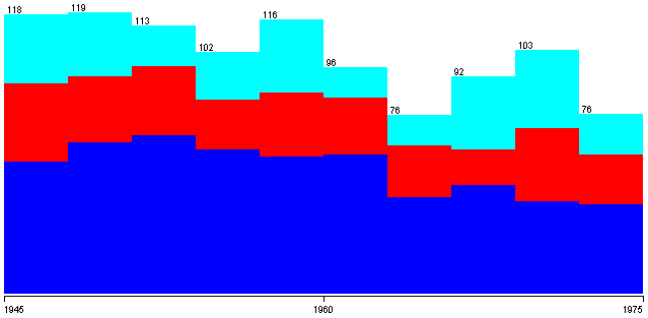

Fig. 2. Distribution of the stories count from 1945 to 1975, drawn with Weka [16]. The X axis represents the year and the Y axis represents the sorties count. The color represents the language text. In blue the Luxembourgish texts, in red the French texts and in pastel the German texts.

Fig. 3. A tag cloud showing the most frequent terms appearing in the French stories of the corpus (*http://tagcrowd.com*).

The analysis of the whole corpus shows that the preponderance of industry and travel is confirmed, with approximately 40% of the stories (Table V). When we examine in details the texts, we can see they are related to various topics such as the development of the Luxembourg-Findel airport as well as the dawn of the siderurgical industry at Belval (Esch-sur-Alzette). Recreational activities are also told (i.e. the official opening of the Capitol movie theater in 1947 - it no longer exists today). Undoubtedly, these facts had a certain influence on people who lived in Luxembourg at this time.

TABLE V
TOP CATEGORIES OF FRENCH STORIES. THE NATURAL LANGUAGE UNDERSTANDING API [15] USES A MODIFIED VERSION OF THE IAB WELL-KNOWN TAXONOMY.

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

VI. Conclusion and perspectives

We have proposed in this work a method to create a multilingual corpus of narratives related to Luxembourg during the 1945-1975 period. The corpus construction was realized by using various text mining techniques like languages detection, syntax and grammar checking and Named Entity Recognition. In practice, a Java standalone tool was implemented to extract data from two well-known data sources like Wikipedia and Wikidata. Moreover, this tool relies on robust embedded components like LanguageTool [13] and remote services like Natural Language Understanding API [15].

As a result, this corpus will be used in a software that aims at helping elderly people to share their personal narratives.

Firstly, the corpus will be visualized in a user interface that propose both a timeline and a map to navigate in time and space. A similar work was recently proposed in the 'Wallonia Time Machine' to browse Belgian narratives during different time periods [5].

Secondly, the prototype will allow the users to find, for a given personal narrative, the corpus narratives that are *temporally*, *geographically* and *semantically* close, independently of the language. This approach will be an extension of the '*Concept based query expansion*' method [17].

Finally, the narratives corpus will be used to generate questions to improve the recall of personal anecdotes (for instance: *do you have a story related to the marriage of Grand-Duc Jean with Princess Joséphine-Charlotte in 1953?*). This approach was recently proposed in the *More Than One Story* card game [6]: with a set of questions, this game asks participants to share their personal experiences. To reach this goal, the software will generate questions by applying Natural Language Generation (NLG), a processing technique that aims at constructing human-readable sentences [14].

Acknowledgments

This work was realized in the context of the LOCALE research project[1], funded by a Core grant from the *Fonds National de la Recherche* (Luxembourg).

References

[1] T. Tamisier, R. McCall, G. Gheorghe, and P. Pinheiro, "Visual analytics for interacting on cultural heritage," in *International Conference on Cooperative Design, Visualization and Engineering*. Springer, 2016, pp. 296–299.

[2] P. Gilles and E. Ziegler, "The "historical luxembourgish bilingual public notices database"," 2013.

[3] R. Schäfer, A. Barbaresi, and F. Bildhauer, "The good, the bad, and the hazy: Design decisions in web corpus construction," in *Proceedings of the 8th Web as Corpus Workshop*, 2013, pp. 7–15.

[4] G. Faaß and K. Eckart, "Sdewac–a corpus of parsable sentences from the web," in *Language processing and knowledge in the Web*. Springer, 2013, pp. 61–68.

[5] J. Strötgen and M. Gertz, "A baseline temporal tagger for all languages." in *EMNLP*, vol. 15, 2015, pp. 541–547.

[6] B. Pouliquen, M. Kimler, R. Steinberger, C. Ignat, T. Oellinger, K. Blackler, F. Fuart, W. Zaghouani, A. Widiger, A.-C. Forslund *et al.*, "Geocoding multilingual texts: Recognition, disambiguation and visualisation," *arXiv preprint cs/0609065*, 2006.

[5]http://wallonia-timemachine.herokuapp.com

[6]http://www.simrishamn.se/sv/kultur_fritid/more-than-one-story/

[7] B. R. Monteiro, C. A. Davis, and F. Fonseca, "A survey on the geographic scope of textual documents," *Computers & Geosciences*, vol. 96, pp. 23–34, 2016.

[8] J. L. Leidner, "Georeferencing: From texts to maps," *The International Encyclopedia of Geography*, 2017.

[9] G. P. Zarri, "Representation and management of complex narrative information," in *Language, Culture, Computation. Computing of the Humanities, Law, and Narratives*. Springer, 2014, pp. 118–137.

[10] A. Spitz, V. Dixit, L. Richter, M. Gertz, and J. Geiß, "State of the union: A data consumer's perspective on wikidata and its properties for the classification and resolution of entities," in *Wiki Workshop at ICWSM*, 2016.

[11] A. Kilgarriff, S. Reddy, J. Pomikálek, and P. Avinesh, "A corpus factory for many languages." in *LREC*, 2010.

[12] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," *Communications of the ACM*, vol. 57, no. 10, pp. 78–85, 2014.

[13] D. Naber, "A rule-based style and grammar checker," 2003.

[14] R. Mitkov, *The Oxford handbook of computational linguistics*. Oxford University Press, 2005.

[15] "Natural Language Understanding API," https://www.ibm.com/watson/developercloud/natural-language-understanding.htm, 2017.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[17] Y. Qiu and H.-P. Frei, "Concept based query expansion," in *SIGIR conference*. ACM, 1993, pp. 160–169.