**Post print Version** - Sébastien Caquard, Emory Shaw & José Alavez (2021): How Distant is Close Enough? Exploring the Toponymic Distortions of Life Story Geographies, GeoHumanities, https://doi.org/10.1080/2373566X.2021.1965898

# How Distant is Close Enough? Exploring the Toponymic Distortions of Life Story Geographies

Sébastien Caquard, Emory Shaw and José Alavez

**Abstract:** Stories are now broadly recognized as important sources of geographic information in different domains of the spatial humanities. The methodologies mobilized to identify these spatial data, however, remain the subject of intense debate. In this paper, we contribute to this debate by focusing on what we can learn from the close reading of stories to improve the quality of distant reading approaches. We do this through an in-depth comparative analysis of how toponyms are used across 10 oral life stories of exiles. Results show that a "distant listening" of the number of country names mentioned in these stories provides an accurate representation of their global geographies. However, the finer-scaled geographies of these stories become highly distorted when counting more local toponyms such as neighbourhoods, cities or regions. This study also reveals that results could be improved by accounting for the distribution and repetition of toponyms throughout these stories. Such insights and their nuances are described in this paper with an aim to help narrow the gap between close and distant reading methodologies.

**Keywords**: story mapping; oral history; literary cartography; close reading; distant reading

## 1. Introduction

Stories and places are interrelated, and any in-depth study of one requires the serious consideration of the other. Approaching a story through the places described in it has become a major form of literary analysis and can offer intriguing spatial perspectives on narratives. Conversely, approaching a place through the stories that are associated with it allows for a more thorough understanding of that place, providing insight into the critical intersection of personal, cultural and socio-political dimensions. While stories and places are intertwined, the identification and characterization of places in stories is not always as simple as it might seem. In certain types of stories that are structured around precise geographical information, such as travel diaries, detailed testimonials of hyper-localized events, or geographically-focussed interviews, places may be easier to identify and circumscribe (see Kwan 2008; Pearce 2008; Watts 2010; Mennis, Mason, and Cao 2013), but for most narratives, meaningful spatial information remains buried deep in the story.

There is no unified methodology or "even common perspective" to identify meaningful places in stories (Cooper, Donaldson, and Murrieta-Flores 2016, 8). Methodologies range from very involved and close readings of individual stories to identify the locations that structure these stories, to much more distant reading approaches that often consist of the systematic identification of locations through broadly recognized toponyms (i.e. proper noun place names). This methodological spectrum reflects deep intellectual debates within the humanities between close-reading analyses that aim to identify the "luminous detail" that "unlocks the shared logic of cultural discourses" and the more distant reading approaches that become "a field of pure information that should be managed by a computer" (Kopec 2016, 330). The latter has gained momentum since the end of the 20th century following the work of literary scholars involved in studying large corpora of novels (Jockers 2013), combined with major progress made in computational linguistics and natural language processing (Piper 2018).

However, close readings have a historical legacy across disciplines such as geography, anthropology, sociology, history, and literary studies (Gallop 2007; Love 2013; Bode 2017) and seems to have re-emerged more recently in the humanities as a way to resist the databasification of literary studies (Devereux 2012; Kopec 2016). This "databasification" relies extensively on identifying and counting toponyms mentioned in stories (Gritta et al. 2018) - also known as the "geographic investment" of a text (Wilkens 2013) - but it remains unclear to what extent such bold calculations reflect or distort the geographies embedded in these stories.

In this paper we propose to address this issue through an in-depth analysis of the geographies of 10 life stories of exiled persons living in Canada. In this case, the life story is a particular form of narrative that is expressed orally, co-constructed between a storyteller and an interviewer, and is neither edited nor scripted. Just like any other narrative forms, life stories can be structured geographically around places are expressed in multiple ways. The paper begins with a review of existing approaches developed across a range of disciplines that focus on identifying and mapping places in stories. We then present two methodologies developed to identify places in oral life stories: analogous to close and distant "readings", we refer to these approaches as a close and distant "listening" respectively. We compare the results obtained with both methods to assess the possibilities and limitations of collecting toponyms with a distant listening for identifying important places in these stories. In the final section we outline a series of trends in the way toponyms are used by storytellers to talk about places and we discuss the potential of these trends to improve distant listening methods.

## 2. Geolocating stories: from closer to more distant methodologies

A "close reading" can be defined as "the thorough interpretation of a text passage by the determination of central themes and the analysis of their development" (Jänicke et al. 2015, 84). In general, this qualitative approach is used for in-depth studies of small selections of stories to identify general phenomena, themes and develop or confirm theories that transcend each individual case. In other words, a close reading entails the nuanced exploration of complex narrative structures (Trumpener 2009) and provides the required flexibility to identify and characterize elements of importance in a story (Kermode 1980), including spatial aspects (Murrieta-Flores and Howell 2017). One major issue with this approach is that it is time-consuming and highly dependent on the goals and profile of the interpreter. It is therefore impossible to reproduce and apply to a large batch of stories in a systematic way.

Whereas such close reading approaches usually involve a deep reading of a small selection of stories, distant reading methodologies entail a more superficial analysis of a larger number of documents. An underlying assumption that supports distant reading in narrative studies is that the availability of "huge data sets means that many areas of research are no longer dependent upon controlled, artificial experiments or upon observations derived from data sampling" (Jockers 2013, 7). Entire corpora can then be studied through a systematic and comprehensive identification of specific words and word groups which can serve to "generate an abstract view by shifting from observing textual content to visualizing global features of a single or of multiple text(s)" (Jänicke et al. 2015, 2). From a spatial perspective, the use of distant reading approaches is often based on the assumption that a narrative is filled with toponyms that can be identified more or less automatically (Cooper, Donaldson, and Murrieta-Flores 2016).

From a close reading perspective, the main objective is to attentively and accurately identify meaningful spatiotemporal elements in a story, whereas from a distant reading perspective, the principal goal is to identify places as efficiently and systematically as possible (Gregory and Cooper 2009). Not surprisingly, distant reading approaches have been gaining momentum due in part to the historic surge in stories available in digital formats: from video testimonies to digitized books to social media posts. This has attracted the attention of computer scientists, who have worked collaboratively with linguists to develop corpus linguistics techniques aided by natural language processing (NLP) methods to automate such processes (Murrieta-flores and Howell 2017; Moncla et al. 2019). Whereas for close reading approaches, the burden of data collection is on the analyst (e.g. literary scholar, historian, geographer), with distant readings it has largely been transferred to linguists and computer scientists.

As advocated by historian Todd Presner through their work with audiovisual testimonies of the Holocaust, distant readings offer several advantages including the fact that "computational analysis can provide insights and ethical perspectives that human listening cannot precisely by the way in which it allows a kind of 'distant listening' based on the whole of the archive rather than a selection of representative or even canonical testimonies" (Presner 2016, 194). Although this is a compelling argument in favor of distant reading methods, it should be nuanced by the fact that even an entire corpus can only represent a very small and somewhat biased fraction of all the human experiences of a collective event. Furthermore, not all stories are prone to equal treatment by any given algorithm. For instance, the frequency of identifiable spatial cues such as toponyms can vary dramatically between stories (even within a coherent corpus): certain storytellers mention way more toponyms than others, which can give more spatial weight to their stories in comparison to those with a lower frequency of place name references. These nuances aside, Presner's argument emphasizes a key aspect of distant readings: all stories of a corpus are treated equally, even if they are not fully equal in front of an algorithm.

Using Moretti's argument that distance is a condition of knowledge (Moretti 2000; 2013), Presner also argues that distant listening can help reveal "structures, patterns, and trends that are not discernable when the focus remains on just a handful of close readings of individual texts" (Presner 2016, 198). This argument is valid on the condition that the distant reading process provides a transparently summative picture that is not too distorted nor too dissociated from its original source. Indeed, as pointed out by Bode (2017), distant readings have been criticized for their reductionism and for their inappropriateness for identifying the nuances and complexities of human experience (Trumpener 2009), as well as for their inadequacy at considering the historical context of a story's production and access (Frow 2008). In other words, distant reading approaches have been criticized for their lack of consideration for intra- and extra- contextual elements in narrative analysis. Beyond this lack of consideration and besides the progress made in natural language processing to reduce textual misinterpretation (Gritta et al. 2017), what remains unclear is how much the multiple short cuts associated to any distant listening approach may contribute to revealing highly distorted geographic structures, patterns and trends.

## 3. Close listening and distant listening[1] methodologies

### *Overview of the mapping life stories project*

Between 2007 and 2012, researchers from Concordia University's Centre for Oral History and Digital Storytelling (COHDS) collected over 500 life stories from refugees and exiles now living in Canada as part of the Montreal Life Stories Project (High 2014). Each audiovisual recording, ranging from 30 minutes to 13 hours, involved the telling of one's personal life story and was facilitated by one or two interviewers. While there were indeed subjects of interest to the project, which led to prompts related to family, education, childhood, as well as one's experience of violence, displacement and exile, the focus remained on "knowing with" rather than "knowing about" (High 2014). Thus, despite such guidance, the subject matter and overall flow of discourse often depended on the storyteller. These stories were not collected to be mapped, but some of them included rich and detailed descriptions of events associated to places. Ten of these were selected for reasons related to their richness in geographical content: five by individuals of Rwandan origin and five of Haitian origin.

Transcription was considered but not performed on these ten stories. Yet while it is true that in oral history, working with the rawest medium (i.e. audiovisual form) is essential to avoid the distancing, interfering, and "rigid linearity" of the transcription process (High and Sworn 2009), working with raw audiovisual material severely limits the potential for any automated parsing and analysis because most such tools work only with text. Acquiring text from such material is however extremely time intensive when done accurately, since it requires a human transcriber, or wildly inaccurate, in this case, if done using automated methods. Indeed, despite promising and increasingly commonplace voice-transcription technology, the accuracy, efficiency and accessibility of transcription technologies remained insufficient until recently (Gregory et al. 2015). This is especially true when the primary goal is to identify place names in a language other than English, with the added complication of less-documented regional accents and extensive usage of vernacular, local or historically overwritten toponyms (Hannun 2017; Tatman 2017), as was the case in this project.

### *Close listening methodology*

An elaborate database was constructed through a close listening of each of the ten life stories. Our approach was inspired by methodologies such as the one developed by Piatti and colleagues (2009) to map novels and by Caquard and Fiset (2014) to transform movies into mappable "geographic bits and pieces", but was also grounded in our initial readings of the interviews. Segmentation criteria, as well as the attributes recorded for each segment, were refined using a 5-hour long Rwandan life story. Once the methodology was considered stable, it was applied to the other nine stories.

Segmenting a narrative for analysis is not new. Barthes proposed the need to dissect discourse for fragments in order to reconstruct a coherent whole (Barthes, 1964) along with a set of

[1] The terms "close listening" and "distant listening" (as opposed to "close" and "distant reading") are used henceforth since they better describe our methodology which used audiovisual material and explored measures such as discourse time, as opposed to measures more appropriate for textual analyses such as character distance.

units or levels that made narrative structurally analysable (Barthes and Duisit 1975). Although the sets of units and levels they propose might be relevant for structural narrative analysis, it appeared less appropriate for identifying mappable spatial units. Gérard Genette (1972) offered an approach involving what he called "narrative segments" that are delimited based on important temporal and/or spatial breaks in the story. Yet as pointed out by Juliette Morel (2018), the definition of "important break" remains vague and persistently subjective. Indeed, when structuring a story in a spatial database, a central challenge remains to determine where the limits of each unit should be situated; in other words, what are the segments of the story that are coherent as well as relevant and important to map?

To segment our stories into coherent "story units", we decided to delimit units in accordance with specific kinds of change in one's narrative. Change was understood as a shift from one place to another, from one point or period of time to another, and/or from one theme or referent to another. In practice, this generated a broad spectrum of story unit types that ranged from hyper-localized spatiotemporal building-blocks within an interviewee's anecdote, to units that could encompass several minutes of narration about a given event in a specific town, to a general reflection about life in a particular country or even a constellation of them. To use Ryan's (2012) spatial narratological terms, our story units would be analogous to a spatial frame, a setting, or even a much broader story space.

*Table 1. Examples of the variable degrees of spatiotemporality among story units that were generated by analysts during the close listening process.*

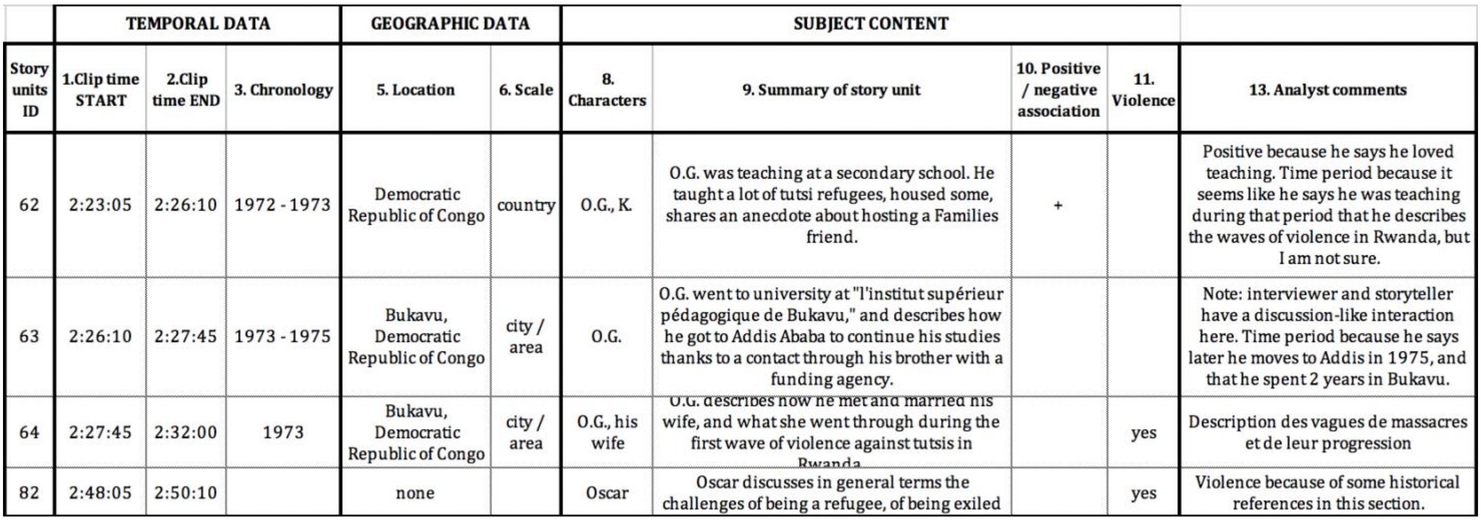
| | | **SPATIALITY** | |
|---|---|---|---|
| | | **HIGH** | **LOW** |
| **TEMPORALITY** | **HIGH** | E.P. describes the itinerary of her escape through the streets and neighbourhoods of Port-au-Prince on a specific day. | E.H. describes how he felt he could no longer run from his Hutu identity since the genocide (*Note: the timespan is clear, and the location is known but not referred to nor important in the content of the story unit).* |
| | **LOW** | J.M. describes the people, landscape and vibe of the city of Marchand Dessalines (*Note: there is no temporal cue here whatsoever: the storyteller describes the city in a general sense having only visited it a handful of times).* | B.N. describes how when people learn history, you don't know what they will take away from it. |

Although this framework was extremely useful to guide the analysts' decisions, in terms of defining story unit boundaries, there were many instances where the story's complexity required the direct interpretation of the analyst, generating more complex story units, such as when more than one place and/or time period and/or subject was linked to a single-story unit. In this case, rather than generating a sequence of small, consecutive story units, a single-story unit encompassed, for example, a comparison between the different places, times and/or subjects. The combination of one or two of these elements was usually grounded by a singular third (for example, a single place could unite a series of topics and times into a single unit, or a single topic could unite a series of places into a unit). A single-story unit could thereby include more than one place, event and/or time period, yet at least one of these three variables remained constant throughout a single unit. As such, story units could be considered as spatial or non-spatial, temporal or non-temporal (Table 1).

All 10 selected interviews were exhaustively divided, into story units that served as the building blocks (i.e. rows) in a close listening database. This database consisted of 14 fields that were identified as both relevant and possible to fill in a consistent manner. These fields included data related to temporality, spatiality, and events along with story unit ID and analyst comments (see Table 2). This last field was designed to record and explain all the choices that were made by the analysts (e.g. interpretations, inferences, omissions, justifications), demonstrating our commitment to making the close listening process as transparent as possible.

Our close listening database was filled by listening carefully to each story's unfolding by two analysts. One analyst handled the first listening and initial data creation, followed by a close listening by a second analyst who validated the choices of the first analyst. In case of disagreement, the team would meet to make the final decision about what to enter into the database.

*Table 2. Example of the database produced with our close listening methodology.*

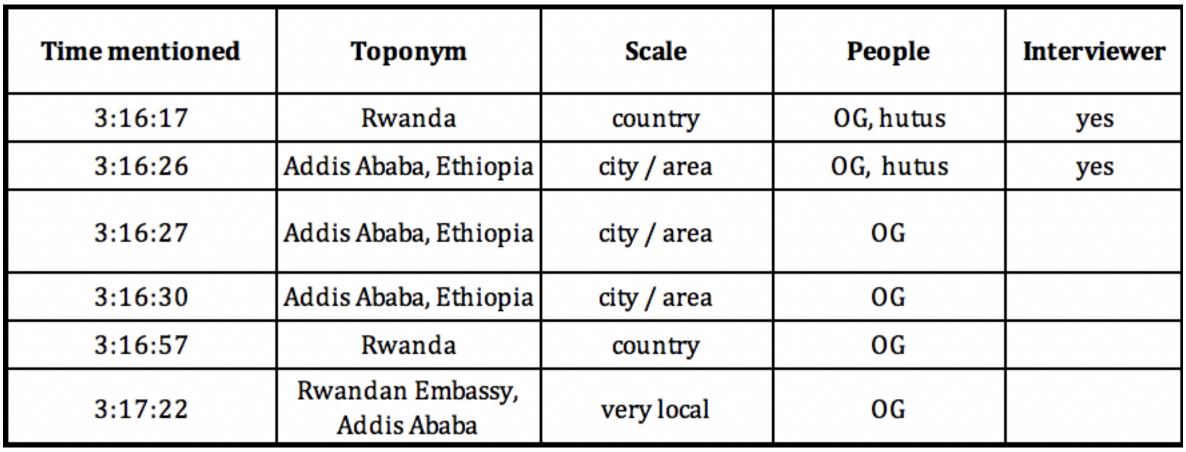


| | TEMPORAL DATA | | | GEOGRAPHIC DATA | | SUBJECT CONTENT | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Story units ID | 1.Clip time START | 2.Clip time END | 3. Chronology | 5. Location | 6. Scale | 8. Characters | 9. Summary of story unit | 10. Positive / negative association | 11. Violence | 13. Analyst comments |
| 62 | 2:23:05 | 2:26:10 | 1972 - 1973 | Democratic Republic of Congo | country | O.G., K. | O.G. was teaching at a secondary school. He taught a lot of tutsi refugees, housed some, shares an anecdote about hosting a Families friend. | + | | Positive because he says he loved teaching. Time period because it seems like he says he was teaching during that period that he describes the waves of violence in Rwanda, but I am not sure. |
| 63 | 2:26:10 | 2:27:45 | 1973 - 1975 | Bukavu, Democratic Republic of Congo | city / area | O.G. | O.G. went to university at "l'institut supérieur pédagogique de Bukavu," and describes how he got to Addis Ababa to continue his studies thanks to a contact through his brother with a funding agency. | | | Note: interviewer and storyteller have a discussion-like interaction here. Time period because he says later he moves to Addis in 1975, and that he spent 2 years in Bukavu. |
| 64 | 2:27:45 | 2:32:00 | 1973 | Bukavu, Democratic Republic of Congo | city / area | O.G., his wife | O.G. describes how he met and married his wife, and what she went through during the first wave of violence against tutsis in Rwanda | | yes | Description des vagues de massacres et de leur progression |
| 82 | 2:48:05 | 2:50:10 | | none | | Oscar | Oscar discusses in general terms the challenges of being a refugee, of being exiled | | yes | Violence because of some historical references in this section. |

### *Distant listening methodology*

The distant listening approach we employed aimed to identify and collect all the toponyms mentioned in each story. Our methodology relied on actively listening for proper noun place names. We listened to each of the 10 stories under study and systematically recorded all discernible toponyms mentioned by either the storyteller or interviewer (see table 3). Like the close listening, this task was carried out by an analyst, followed by a second listening by another analyst for validation. In the spirit of remaining "distant", analysts did not take note of common noun place names such as "town" or "airport", even though they might have been able to deduce the exact location of a "town" or "airport" reference based on the story's context.

*Table 3. Example of the database produced with our distant listening methodology.*



| Time mentioned | Toponym | Scale | People | Interviewer |
|---|---|---|---|---|
| 3:16:17 | Rwanda | country | OG, hutus | yes |
| 3:16:26 | Addis Ababa, Ethiopia | city / area | OG, hutus | yes |
| 3:16:27 | Addis Ababa, Ethiopia | city / area | OG | |
| 3:16:30 | Addis Ababa, Ethiopia | city / area | OG | |
| 3:16:57 | Rwanda | country | OG | |
| 3:17:22 | Rwandan Embassy, Addis Ababa | very local | OG | |

Once the toponyms were identified, we used the Google Places gazetteer as well as Geonames to geocode them. Among the multiple challenges faced during this process was the issue of associating coordinates to historical place names (Heuser et al. 2016). This was especially true for some places in Rwanda, since the country's administrative divisions were redrawn in 2006: long after the five concerned storytellers had left the country. Since these historical place names were not systematically available on Google Places and Geonames, some additional research was necessary to locate them. Overall, less than 10 local toponyms (less than 0.3 % of the total) had to be omitted since no explicit locations for them were found anywhere on the Internet.

## 4. Comparing geodatabases

A series of comparative analyses were carried out to assess the differences between geographic data produced by the distant listening and close listening methodologies, with the latter being used as a reference for evaluating the accuracy of the former. The following subsections describe the methods used and developed to perform these comparisons.

### *Overview of the analytical process*

Throughout the close listening process, our ten stories were divided into 1,839 fragments or "story units" (SU), of which 78% (i.e. 1,434) were considered spatial and therefore associated to one or more places and their respective spatial coordinates (see table 4). These 1,434 spatial story units (SSU) became our main spatial reference for the comparative analysis that followed. SSUs were compared to the 3,266 toponyms identified throughout the distant listening process.

It is worth noting that there were more than twice as many unique geolocated places in the distant listening database (i.e. average of 44 per story) as in the close listening one (i.e. average of 21 per story) (see Table 5). This suggests that overall, about half of the unique toponyms mentioned in each story were not considered important enough during the close listening process to define the spatiality of at least one story unit. In other words, they did not define settings within the narrative. For instance, when a storyteller briefly mentions that his cousin now lives in Belgium, while the toponym is mentioned and therefore appears in the distant listening database, "Belgium" is not considered important enough in this story to justify the creation of its own story unit and so does not appear in the close listening database. Since only half as many toponyms mentioned in these stories were considered important enough to be associated with story settings (i.e. SSU), this raised the question of how to distinguish these toponyms from the more accessory ones to improve the quality of the distant listening process.

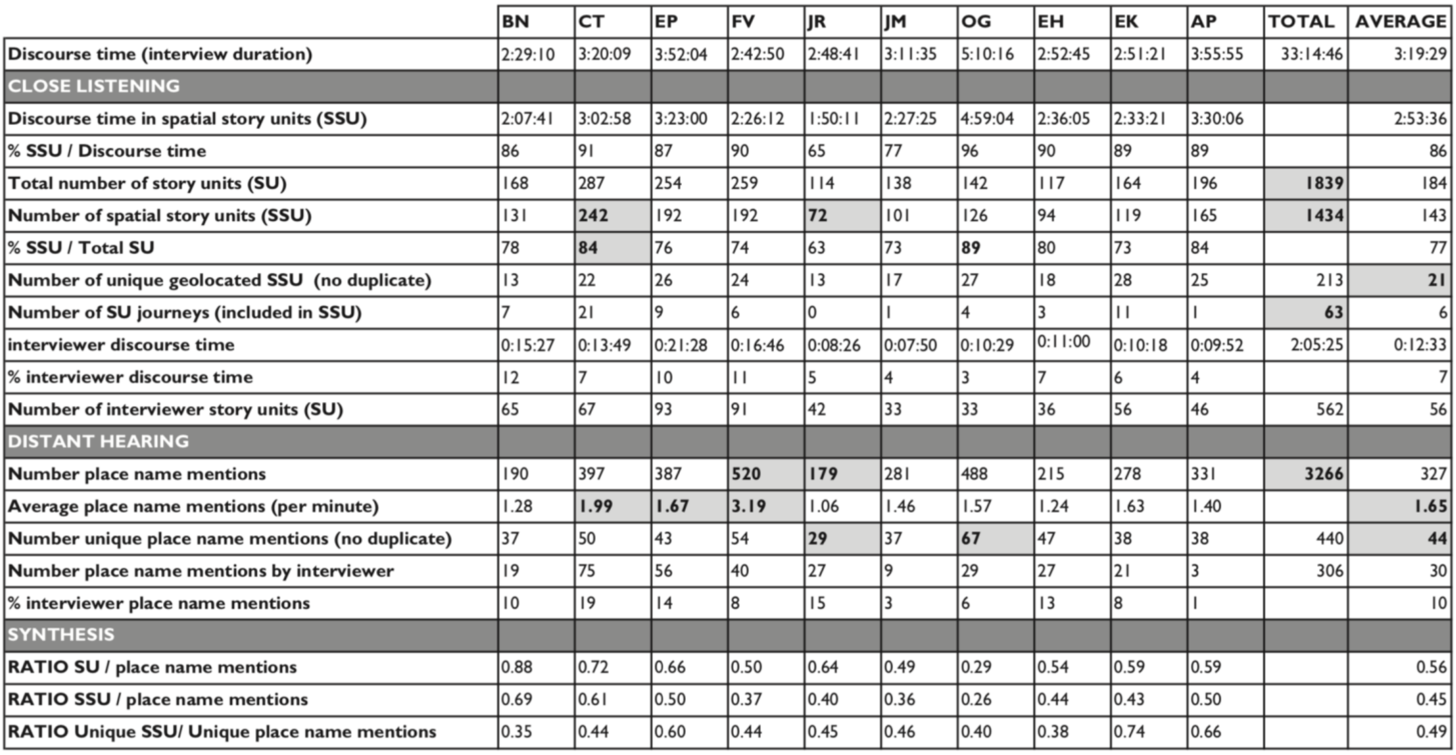
*Table 4. Overview of the content of the two databases.*

| | **Close listening** | **Distant hearing** |
|---|---|---|
| **Unit of measurement** for comparative analysis | **Time** (in minutes and seconds) associated to each SSU | **Frequency** (i.e. Number of times a place name is mentioned) |
| **Data compared** during the analysis | **1 736 min** (from 1 434 SSU) in total | **3 266 place names** mentioned in total |
| Total **entries** for the 10 selected stories | **1,434** Spatial Story Units (SSU) and 63 journeys within a total of 1 839 story units (SU).<br>The number of SSU per story ranged from 74 to 242. | **3,266** place names mentioned (including 306 mentioned by interviewers).<br>The number of place names mentioned per story ranged from 179 to 520. |
| Total number of **unique geolocated places** (i.e. excluding duplicate) | **213** (Note: the number of geolocated places per story range from 13 to 28) | **440** (Note: the number of geolocated places per story range from 29 to 67) |

### *Visual explorations*

To address this question, a first step was then to produce a series of visual comparisons using Atlascine 3.0, an online, open-source mapping application developed to map stories for research purposes (see Caquard and Fiset 2014; Caquard and Dimitrovas 2017). In the context of this comparative analysis, Atlascine was used to visually compare the two databases at scales varying from global to local (see Figure 1) as well as in different parts of the globe (e.g. country of origin versus country of destination) and at various points throughout each story. Atlascine employs geolocated, tree-ring proportional symbols to represent the relative importance of places either because the storyteller spends a large amount of time talking about events associated to these places (close listening), or because their name is mentioned often (distant listening) (Caquard and Fiset 2014). Various insights were gleaned from the comparison of these map pairs. For instance, across all ten stories, each pair appeared similar in general terms, but with some important nuances at different scales. What stood out most was that the smaller the scale (i.e. when zoomed out at a continental scale), the more the two maps of each pair resembled each other. We also noticed that the apparition of the toponyms on the two maps were not always synchronized. This led us to design a series of spatio-temporal graphs to compare the synchronization of both databases throughout each story (see Figure 2).

*Table 5. Quantitative summary of each of the stories under study (note: the results mentioned in the text appear in bold and grey in the table).*



| | BN | CT | EP | FV | JR | JM | OG | EH | EK | AP | TOTAL | AVERAGE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Discourse time (interview duration)** | 2:29:10 | 3:20:09 | 3:52:04 | 2:42:50 | 2:48:41 | 3:11:35 | 5:10:16 | 2:52:45 | 2:51:21 | 3:55:55 | 33:14:46 | 3:19:29 |
| **CLOSE LISTENING** | | | | | | | | | | | | |
| **Discourse time in spatial story units (SSU)** | 2:07:41 | 3:02:58 | 3:23:00 | 2:26:12 | 1:50:11 | 2:27:25 | 4:59:04 | 2:36:05 | 2:33:21 | 3:30:06 | | 2:53:36 |
| **% SSU / Discourse time** | 86 | 91 | 87 | 90 | 65 | 77 | 96 | 90 | 89 | 89 | | 86 |
| **Total number of story units (SU)** | 168 | 287 | 254 | 259 | 114 | 138 | 142 | 117 | 164 | 196 | **1839** | 184 |
| **Number of spatial story units (SSU)** | 131 | **242** | 192 | 192 | **72** | 101 | 126 | 94 | 119 | 165 | **1434** | 143 |
| **% SSU / Total SU** | 78 | **84** | 76 | 74 | 63 | 73 | **89** | 80 | 73 | 84 | | 77 |
| **Number of unique geolocated SSU (no duplicate)** | 13 | 22 | 26 | 24 | 13 | 17 | 27 | 18 | 28 | 25 | 213 | **21** |
| **Number of SU journeys (included in SSU)** | 7 | 21 | 9 | 6 | 0 | 1 | 4 | 3 | 11 | 1 | **63** | 6 |
| **interviewer discourse time** | 0:15:27 | 0:13:49 | 0:21:28 | 0:16:46 | 0:08:26 | 0:07:50 | 0:10:29 | 0:11:00 | 0:10:18 | 0:09:52 | 2:05:25 | 0:12:33 |
| **% interviewer discourse time** | 12 | 7 | 10 | 11 | 5 | 4 | 3 | 7 | 6 | 4 | | 7 |
| **Number of interviewer story units (SU)** | 65 | 67 | 93 | 91 | 42 | 33 | 33 | 36 | 56 | 46 | 562 | 56 |
| **DISTANT HEARING** | | | | | | | | | | | | |
| **Number place name mentions** | 190 | 397 | 387 | **520** | **179** | 281 | 488 | 215 | 278 | 331 | **3266** | 327 |
| **Average place name mentions (per minute)** | 1.28 | **1.99** | **1.67** | **3.19** | 1.06 | 1.46 | 1.57 | 1.24 | 1.63 | 1.40 | | **1.65** |
| **Number unique place name mentions (no duplicate)** | 37 | 50 | 43 | 54 | **29** | 37 | **67** | 47 | 38 | 38 | 440 | **44** |
| **Number place name mentions by interviewer** | 19 | 75 | 56 | 40 | 27 | 9 | 29 | 27 | 21 | 3 | 306 | 30 |
| **% interviewer place name mentions** | 10 | 19 | 14 | 8 | 15 | 3 | 6 | 13 | 8 | 1 | | 10 |
| **SYNTHESIS** | | | | | | | | | | | | |
| **RATIO SU / place name mentions** | 0.88 | 0.72 | 0.66 | 0.50 | 0.64 | 0.49 | 0.29 | 0.54 | 0.59 | 0.59 | | 0.56 |
| **RATIO SSU / place name mentions** | 0.69 | 0.61 | 0.50 | 0.37 | 0.40 | 0.36 | 0.26 | 0.44 | 0.43 | 0.50 | | 0.45 |
| **RATIO Unique SSU/ Unique place name mentions** | 0.35 | 0.44 | 0.60 | 0.44 | 0.45 | 0.46 | 0.40 | 0.38 | 0.74 | 0.66 | | 0.49 |

At first glance, the spatio-temporal graphs show that countries and cities are the two most common spatial scales in these stories; local, very local and non-spatial references are less common, whereas continents, sub-national regions and journeys were marginal. That said, there are some clear exceptions. For instance, most of the places mentioned and associated with CT's story are at the local scales while many segments of JR and JM stories were considered as non-spatial. Some stories remain at the same scale consistently from beginning to end (e.g. AP at the city level) while others constantly shift between scales and are brimming with spatial references overall (e.g. FV). In terms of how these different spatial scales are mobilized throughout the development of these stories, there is no clear pattern either. A storyteller can start talking about a country (e.g. OG), a city (e.g. CT), or an even finer-scale place (e.g. EP, EH) and finish their story at the same scale (e.g. CT) or at a completely different one (e.g. EP, EH). Overall, these graphs illustrate the diversity of spatial scales mobilized within each story and the complexity of identifying overarching spatial patterns between them; it led us to mobilize statistical methods to try to identify finer and more subtle relationships between databases.

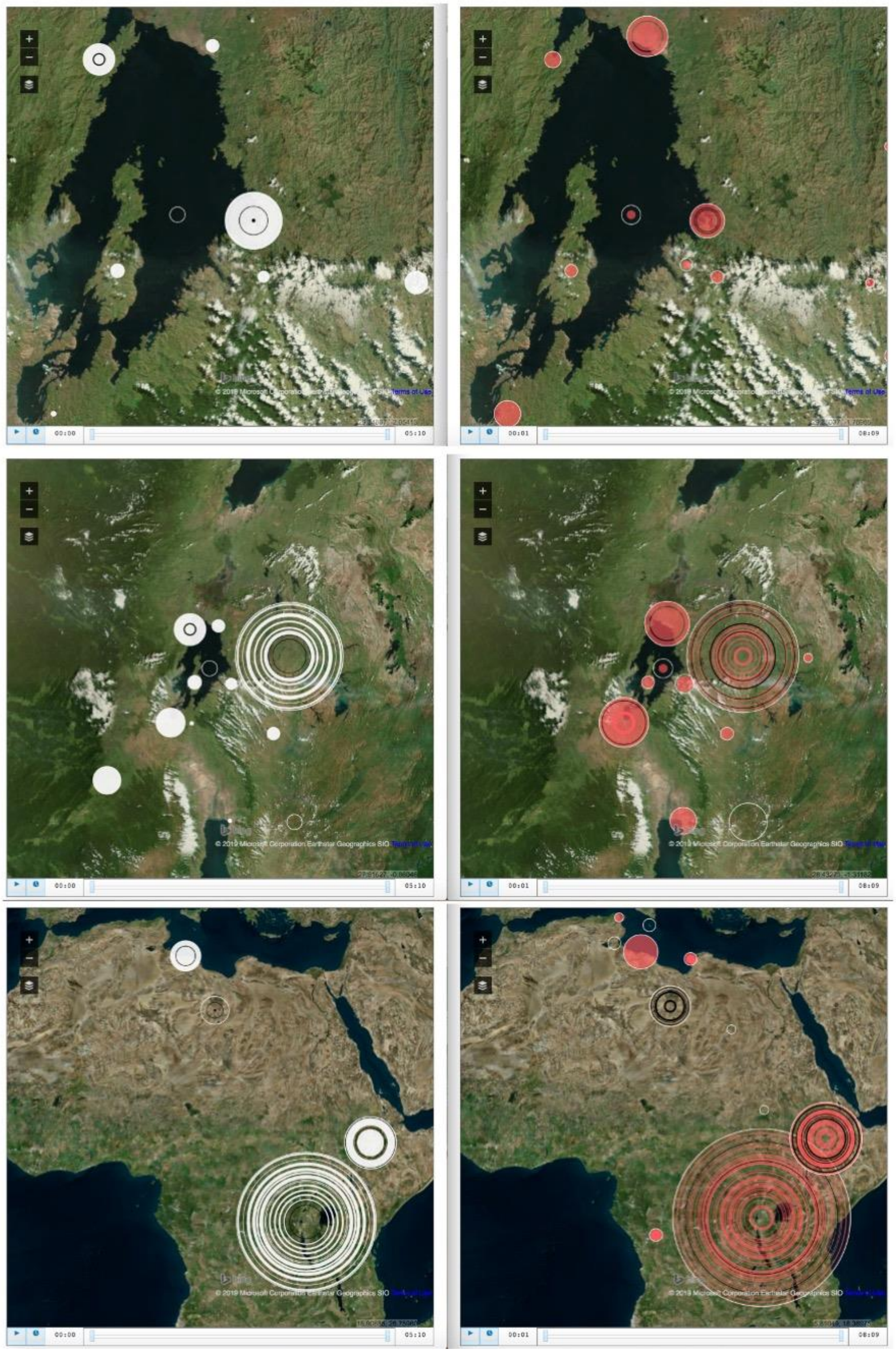


*Figure 1. A visual comparison of both databases at three different scales for the life story of OG. On the left, the proportional symbols represent the amount of time associated to spatial story units (SSU, from the "close listening"), whereas on the right they represent the frequency of toponyms mentioned (place mentions, from the "distant listening"). The larger the ring, the more time the storyteller spent talking about the location (left) or the more frequently the place name was mentioned (right). Black tree rings represent the SSUs or mentioned place names associated with the interviewer.*

We then completed these visual comparisons with a series of simple regression analysis that were performed within and between each database at multiple scales for different regions and story segments. These results combined with our deep engagement with each story through close listenings, and the study of the different maps and spatio-temporal graphs produced, led us to identify a series of trends and provide suggestions that might help inform the use of toponyms in distant listening/reading approaches of stories.

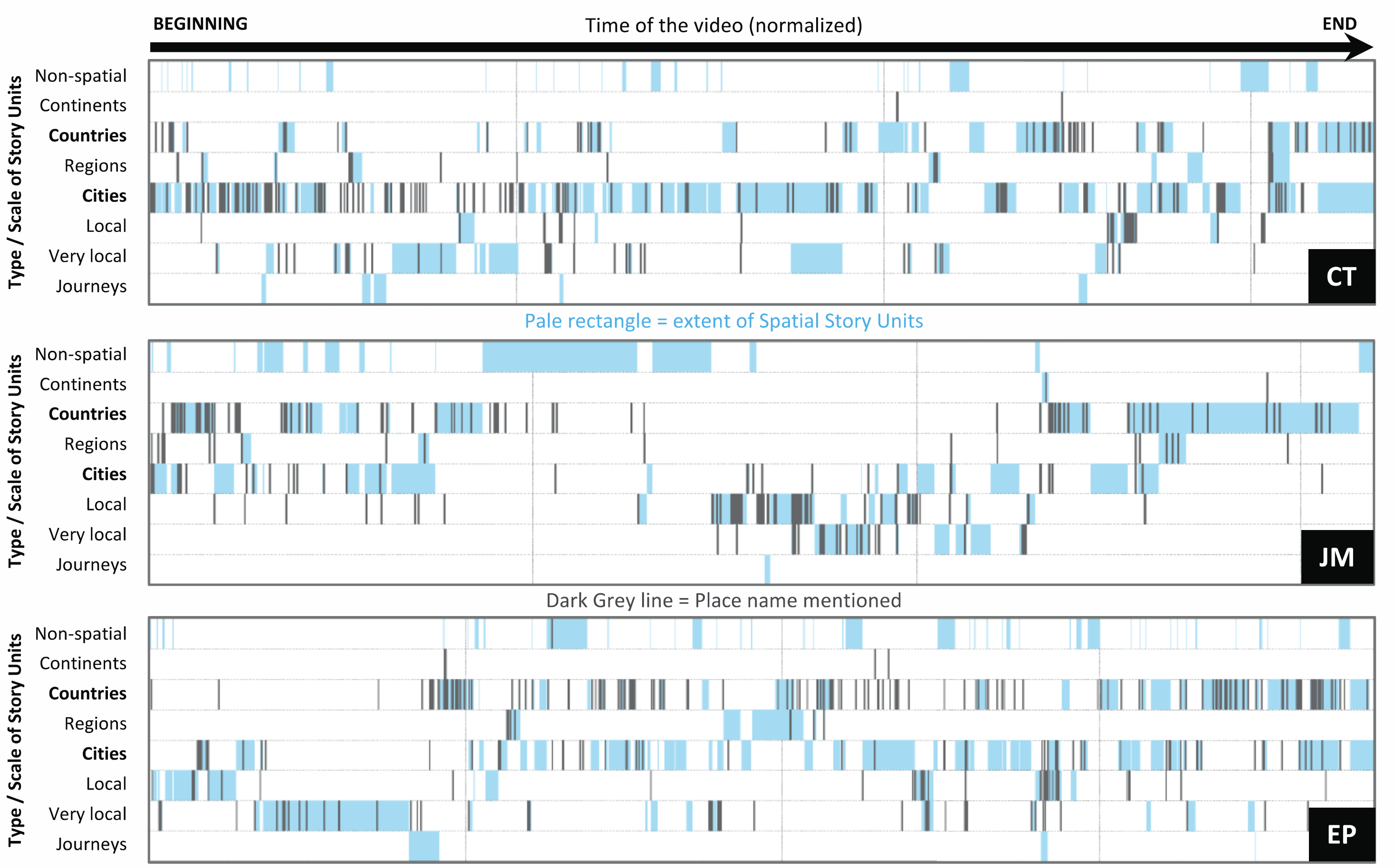


*Figure 2. Examples of graphs showing both the extents of story units in discourse time (light blue horizontal spans) and the place name mentions (dark grey vertical lines) aggregated by scale for each of stories. Note that the length of each story has been normalized here. All the ten graphs and their data can be consulted interactively online (https://github.com/maphouse/lifestories).*

## 5. Beyond counting toponyms

### *Country scale toponym counts are accurate estimates of story geographies*

To explore how geographical scales might influence a distant reading's quality, we performed simple regression analysis in which we compared, for each story and at different scales (i.e. national, regional, local, very local), the number of times each toponym was mentioned in the distant listening database with the amount of discourse time associated with it in the close listening database. The relationship was particularly significant at the national scale (see Figure 3). On average, 13 different country names were mentioned per story (ranging from 6 to 27).

Across all stories, the coefficient of determination ($R^2$) obtained when comparing country names between both datasets was high, ranging from 0.80 to 0.97; depending on the story, between 80 % and 97% of the amount of time associated with each country in our spatial story units (SSUs) (i.e. close listening) could be estimated by simply counting the number of times each country name was mentioned by a storyteller using the distant listening methodology. This suggests that counting the number of country name utterances in a story could provide an accurate, global estimate of the relative importance of countries in a story.

It is important to mention that these results did not improve by taking into account all the sub-national place names and aggregating them at the national level (e.g. when cities like "Port-au-Prince" and "Jérémie" were counted as "Haiti" since they are cities located in this country). Counting basic country names was as good as counting all the toponyms and aggregating them at the country level to assess the global geography of these stories.

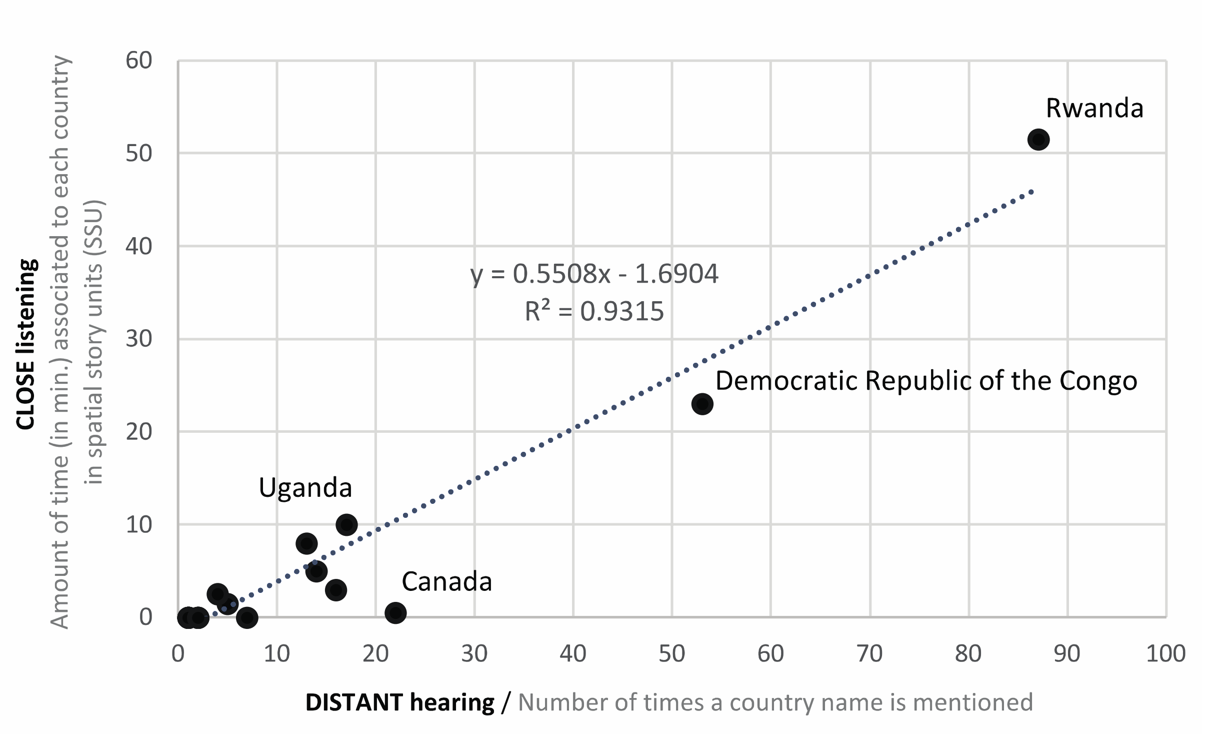


*Figure 3. Scatter plot representing the relationship between the number of mentions and discourse time for the 20 country names found in OG's life story (note: several points with the values (1, 0) and (2, 0) overlap). As can be interpreted by the R-square value, 0.93 of the variation in y can be explained by x.*

Country names were extensively used by storytellers, representing an average of 50% of all toponyms mentioned (between 18% and 69% depending on the story). One of the reasons for this high average is due to the use of country names to serve as a geographical proxy, as illustrated in AP's story. AP, who is originally from Jérémie, a small city where she spent most of her time in Haiti, mentions Jérémie only 3 times in her entire story, but mentions Haiti 115 times. We observed that AP often used "Haiti" to describe events that actually happened in "Jérémie", as is made clear by the fact that the latter is associated with a total of 36 minutes of discourse time in our close listening database (for only 3 mentions) versus 46 minutes for "Haiti" (for 115 mentions). This extensive use of country names to refer to more local places explains in part the lack of correlation between both databases at finer, sub-national scales (i.e. provinces, large cities, towns, villages or neighborhoods).

### *Local scale toponym counts are inaccurate estimates of story geographies*

Despite multiple comparisons done at various scales with different subsets, we identified no clear correlations between close and distant reading datasets at more local scales (addresses, neighborhoods, cities or local regions). Subnational toponyms gleaned by distant listenings provided highly distorted geographical representations of discourse time for the majority of stories under study.

However, in three out of the ten cases, there was a strong correlation between local mentions (i.e. local regions, cities, and towns) and discourse time in cities (i.e. R2 = 0.87 for EK; 0.94 for JR; 0.96 for FV). This suggests that for certain stories, distant readings of some local toponyms might provide a good estimate of the relative importance of these local places in the narrative. But, despite looking at these stories through multiple angles - such as the size of the cities and towns mentioned, their geographical distribution, their position within the story - we were not able to determine the conditions that might explain these results.

What we noticed though is that storytellers tend to use more local toponyms to refer to places in their country of destination (i.e. Canada), than in their countries of origin (i.e. Haiti and Rwanda). Overall, subnational toponyms represent 78% of the Canadian place names (i.e. "Canada" covers the remaining 22%) in comparison to only 37% and 43% of all the Haitian and Rwandan subnational toponyms respectively. This might be due to a combination of factors, such as the context of the interview, the implicit or explicit interests of the interviewer or the storyteller's understanding of the interviewer's background. For example, 6 out of the 10 main interviewers were youth of second generation from the community of the exiled storyteller. These interviewers had possibly not spent as much – if any – time in the countries of origin of their elders and might have more in common with spatial references in Canada and Montreal in particular. This raises the question of the interviewer's influence on such inherently dialogical story geographies.

### *Interviewer utterances influence certain story geographies*

Many stories displayed an influence of the interviewer on the life story's geography as told by the interviewee. The three stories with the highest average of place names mentioned (FV, CT, EP, see Table 5) were from three interviews conducted at least partially by a graduate student in geography with a particular interest in questions of identity and place (Roux 2009). This particular interest is reflected in the total number of toponyms mentioned by this interviewer, which was higher than any of the other interviewers. That said, our observations showed that a storyteller can still be spatially explicit without the influence of an interviewer. For instance, whereas AP's interviewer only mentioned 3 place names throughout their entire interview (way below the average of 30 for the 10 interviews), AP mentioned 331 place names in her story (slightly above the average of 327 for all stories) (see Table 5). In other words, while an interviewer can certainly influence the quantity and precision of place names mentioned, there is no clear correlation between the number of toponyms mentioned by the interviewer and the overall number of toponyms in the story.

A closer analysis of the interaction between the interviewer and the storyteller showed that, for all interviews, when a toponym was mentioned by the interviewer and then repeated by the storyteller, it was usually considered as significant during the close listening process and associated with an SSU. This was observable at all scales, but was especially true for the country-scale. Conversely, when a toponym was mentioned by the interviewer, but not followed by a mention by the storyteller, this place was usually not registered as an SSU. In brief, this might suggest that a toponym mentioned by an interviewer is more likely to be geographically salient when confirmed in the form of a repetition by the storyteller. On the other hand, a toponym mentioned by the interviewer but which is not repeated afterward by the interviewee is unlikely to be considered important at this particular moment of the story.

### *Temporally-clustered toponyms often indicate important places*

When toponyms are repeated over short periods of time, they are likely associated with a SSU, and are therefore indicative of significant places in discourse. For example, this is apparent in OG's story, in which the cities of Gisenyi and Goma are mentioned almost as many times each throughout the story (8 and 10 times respectively), but only Gisenyi was identified (once) as a SSU. This SSU coincides with 4 consecutive mentions of it, whereas the other 4 mentions of Gisenyi are dispersed throughout the story just like the 10 mentions of Goma. The association between temporally clustered toponyms and discourse time extends across scales too. These observations suggest that temporal clusters of same toponyms are a good indicator of a story setting at this location and at that moment in a story. The probability for a cluster to be associated to a SSU appeared to increase with the number of times a toponym was repeated in a short period of time beyond a certain threshold, but decreased with the close presence in discourse time of different toponyms.

Inversely, the lower the density of toponyms in a segment of a story, the less likely this toponym was to be associated to a SSU at that particular moment. Isolated toponyms often appear in reflective discourse; a type of discourse we observed across these stories that can involve conceptual or universal reflections about life, politics or social relationships. They are usually associated with non-spatial story units and therefore not relevant to map. For instance, at a point when EP mentions "Haiti", she talks about the general perception that communism is associated to intellectuals all over the world, including in Haiti. This mobilization of toponyms in reflective discourse is uncommon. Indeed, whereas the average number of toponyms mentioned per minute for the 10 stories is 1.6, it drops to 0.4 for the non-spatial SU which are largely associated to reflective discourse. Looking at whether toponyms are isolated from other toponyms in discourse might help to disregard those not associated to spatial events.

### *The spatial and non-spatial significance of flyover toponyms*

We observed a way to gauge the importance of toponyms as anchors for spatial discourse by considering where they are positioned geographically in relation to each other. In other words, once geocoded, what is the geographical distance between subsequently mentioned places? In general, toponyms mentioned in geographic isolation from those before and after them were unlikely to have much importance in defining a story's geography.

A frequently noted occurrence in this study was that when someone was discussing things generally to illustrate a point, they would sometimes use toponyms that were spatially scattered. Not only were place mentions sparser and more infrequent in such reflective discourse than during more descriptive and narrative discourse, but they tended to be accompanied by large geographical distances between them. For example, in the case of JR, the double mention of "France" preceded by a mention of "Iceland", and followed directly by a mention of "Italy" characterize the use of country names not as locations, but as referents for illustrating a point. In this segment, JR was not talking about events related to these places, but about personal values as acquired by his parents as part of their Rwandan culture. Referred to here as "flyover toponyms", identifying these might help identify discourse fragments that are less likely to be significant in spatial narrative terms. However, we found that flyover toponyms could be easily mistaken for journeys, which are key geographic features in narratives of exile.

Journeys tend to feature two or more toponyms which include an origin and a destination, with potentially numerous places in between. Indeed, within journeys, place names are often told in a row, as seen in BN's life story: "Germany", "Canada", "Dorval" and then "Ste-Anne-de-la-Perade". Although journeys can be difficult to differentiate from flyovers when described in this way, they usually include a higher density of toponyms on average (2.7 per minute). This density combined with other linguistic clues (e.g. travel related verbs between toponyms, narrative style as determined by past participles, etc.) might help to properly differentiate between geographically-salient journeys and flyover toponyms using distant readings.

### *Identifying and linking place proxies*

Expanding the analysis beyond proper noun place names or toponyms may help improve the quality of the spatial assessment with distant listening, but here again, the relevant criteria vary widely between stories. As mentioned previously, certain country names can serve as proxies to talk about more local places, just like common noun place names or people. For instance, Karama and Rukondo are important places in EH's story (7 and 6 SSU respectively), since they are the homes of his grandparents (Karama) and parents (Rukondo) as well as where he spent most of his childhood. But these two toponyms are barely mentioned throughout his story (twice and once respectively). In fact, they are mostly referenced through a combination of more vague spatial terms (e.g. "Rwanda") and people associated to them (e.g. "chez ma grand-mère", "chez mes parents"). In a different story, OG describes his childhood extensively, but mentions the place where he spent his childhood (i.e. Kibuye) only very late in the story, so the connection between his childhood and Kibuye can only be made retroactively and, in our case, through a close listening.

These few examples illustrate the subtle ways in which multiple toponyms are mobilized in combination with all sorts of terms by storytellers to talk about events and journeys, to describe places and memories as well as to share ideas and feelings in ways that are unique to each storyteller and to each story. The way toponyms are used in a story can reveal quite a bit about the meaning of a city, a neighborhood, a mountain, or a country, but a large part of this meaning remains hidden somewhere in the gap between close and distant reading or listening. Through the different trends identified in this paper, we hope to contribute to the narrowing of this gap.

## 6. Conclusion

To what extent can the geography of a life story be circumscribed through a distant reading of the toponyms mentioned in it? In this research, which was based on a distant and close listening of ten oral life stories of exiles, we found that counting the number of times country-scale place names were mentioned was a reliable and accurate way to represent the overall geography of a life story. However, this finding did not apply to more local place names (e.g. sub-national regions, cities, neighborhoods), whose global counts can provide a distorted representation of the relative importance of these local places within a story. Such geographical distortions could be reduced by considering certain nuances which were identified throughout this paper, such as how toponyms isolated in segments with lower toponym densities can be disregarded since they are often associated with non-geographic, reflective discourse. Conversely, toponyms that are repeated and clustered often refer to places of importance, especially when repeated by both discussants in an interview context. Clusters of different toponyms are more complex to interpret, since they can be used in reflective discourse to illustrate points not directly relevant to the overall narrative, or in more geographical descriptions of events that connect different places such as journeys. These observations point to the importance of evaluating the geographical salience of toponyms across a story based on how those toponyms are distributed: their sparseness or density, as well as their repetition and diversity, but the actual geographical distances between geocoded toponyms can also be a factor. These vary immensely depending on a storyteller, their recollection of events, how they wish to reveal themselves to a given audience, interactions with an interviewer, and their respective goals. A life story delivered as a detailed testimony may offer a greater level of geographical detail, whereas a life story that involves more self-reflection and personal discovery may be less attentive to providing geographical detail. All stories, even within the same corpus, are not equal in the face of a distant reading protocol.

The quality of distant readings is likely to improve along with the application of methods driven by artificial intelligence (Gritta et al. 2018), and it is our hope that the observations made here could contribute to these applications for the geographical analysis of stories and discourse in general. The close reading of a selection of stories identified as representative of the main types of geographical discourses within a given corpus will likely remain fundamental to properly training distant reading algorithms. Such tools could help in identifying important places in stories beyond merely stating how often they are mentioned. Doing so would help researchers understand how places are named (or not named) by storytellers, what they tell us about individual and collective geographies embedded in stories and about experiences and memories of place. Closing the gap between close and distant reading methodologies is not just about improving the efficiency and precision of story analysis; it is about expanding the way we envision the multiple geographies embedded in stories and, in turn, what these stories can bring to our understanding of places.

## Acknowledgements

We would like to thank all the individuals mentioned in this paper who made this project possible by sharing their life stories. This project was supported by The Social Sciences and Humanities Research Council of Canada (SSHRC) and CANARIE Canada.

## References

Barthes, R. 1964. Rhétorique de l'image. Communications, 4(1): 40-51.

Barthes, R., and L. Duisit. 1975. An introduction to the structural analysis of narrative. New literary history 6(2): 237-272.

Bode, K. 2017. The equivalence of “close” and “distant” reading; or, toward a new object for data-rich literary history. Modern Language Quarterly 78(1): 77-106. doi: 10.1215/00267929-3699787

Caquard S., and J.P. Fiset. 2014. How Can We Map Stories? A Cybercartographic Application for Narrative Cartography. The Journal of Maps 10(1): 18-25 doi: 10.1080/17445647.2013.847387

Caquard, S., and S. Dimitrovas. 2017. Story Maps & Co. The state of the art of online narrative cartography. Mappemonde 121: 1-31 (http://mappemonde.mgm.fr/121_as1/#englishversion).

Cooper, D., C. Donaldson, and P. Murrieta-Flores. 2016. Introduction: Rethinking Literary Mapping. In Literary Mapping in the Digital Age, ed. D. Cooper, C. Donaldson, and P. Murrieta-Flores, 19-40. 1st. ed. London: Routledge

Devereux, C., 2012. ‘A Kind of Dual Attentiveness’: Close Reading after the New Criticism. In Rereading the New Criticism, ed. Hickman and McIntyre, 218–230. The Ohio State University Press: Columbus.

Frow, J. 2008. Thinking the Novel. New Left Review 49 (Jan Feb): 137-145.

Gallop, J. 2007. The historicization of literary studies and the fate of close reading. Profession : 181-186. DOI: 10.1632/prof.2007.2007.1.181

Genette, G. 1972. Figures III. Paris: Editions du Seuil.

Gregory, I., and D. Cooper. 2009. Thomas Gray, Samuel Taylor Coleridge and geographical information systems: A literary GIS of two Lake District tours. International Journal of Humanities and Arts Computing 3(1-2): 61-84. doi: 10.3366/ijhac.2009.0009

Gregory, I., C. Donaldson, P. Murrieta-Flores, and P. Rayson. 2015. Geoparsing, GIS, and Textual Analysis: Current Developments in Spatial Humanities Research. International Journal of Humanities and Arts Computing 9(1).1-14. doi: 10.3366/ijhac.2015.0135

Gritta, M., M.T. Pilehvar, N. Limsopatham, and N. Collier. 2018. What’s missing in geographical parsing? Lang Resources & Evaluation 52, 603–623. doi: 10.1007/s10579-017-9385-8

Hannun, A. 2017. Speech Recognition Is Not Solved. Available online: https://awni.github.io/speech-recognition/#fn:data_details (Accessed March 4, 2020)

Heuser, R., M. Algee-Hewitt, E. Steiner, and V. Tran. 2016. Mapping the Emotions of London, 1700-1900: A Crowdsourcing Experiment. In Literary Mapping in the Digital Age, ed. D. Cooper, C. Donaldson, P. Murrieta-Flores, 25–46. Abingdon: Routledge. Digital Research in the Arts and Humanities.

High, S. 2014. Oral history at the crossroads: Sharing life stories of survival and displacement. Vancouver: UBC Press.

High, S., and D. Sworn. 2009. After the interview: The interpretive challenges of oral history video indexing. Digital studies/Le champ numérique 1(2) doi: /10.16995/dscn.110

Jänicke, S., G, Franzini, M.F. Cheema, and G. Scheuermann. 2015. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. EuroVis (STARs): 83–103.

Jockers, M.L., 2013. Macroanalysis: Digital methods and literary history. Champaign, IL: University of Illinois Press.

Kermode, F., 1980. Secrets and Narrative Sequence. Critical Inquiry 7, 83–101. doi: 10.1086/448089

Kopec, A., 2016. The Digital Humanities, Inc.: Literary Criticism and the Fate of a Profession. PMLA 131, 324–339. doi: 10.1632/pmla.2016.131.2.324

Kwan, M.P. 2008. From oral histories to visual narratives: Re-presenting the post-September 11 experiences of the Muslim women in the USA. Social & Cultural Geography 9(6): 653-669. doi: 10.1080/14649360802292462

Love, H. 2013. Close reading and thin description. *Public Culture*, *25*(3 (71)), 401–434. doi: https://doi.org/10.1215/08992363-2144688

Mennis, J., M. Mason, and Y. Cao. 2013. Qualitative GIS and the visualization of narrative activity space data. International Journal of Geographical Information Science 27(2): 267-291. doi: 10.1080/13658816.2012.678362

Moncla, L., M. Gaio, T. Joliveau, Y.F.L. Lay, N. Boeglin, and P.O. Mazagol. 2019. Mapping urban fingerprints of odonyms automatically extracted from French novels. International Journal of Geographical Information Science 33, 2477–2497. Doi: 10.1080/13658816.2019.1584804

Morel, J. 2018. Cartographie du Cycle de Nedjma de Kateb Yacine : modélisation spatiale d'un récit littéraire. Les Cahiers d'EMAM. Advance online publication. doi: 10.4000/emam.1444.

Moretti, F. 2000. Conjectures on world literature. New left review, 54-68.

Moretti, F. 2013. Distant reading. London: Verso

Murrieta-Flores, P., and N. Howell. 2017. Towards the Spatial Analysis of Vague and Imaginary Place and Space: Evolving the Spatial Humanities through Medieval Romance. Journal of Map & Geography Libraries 13, 29–57. DOI: 10.1080/15420353.2017.1307302

Piatti, B., H. R. Bär, A.K. Reuschel, L. Hurni, and W. Cartwright. 2009. Mapping literature: Towards a geography of fiction. In Cartography and art, 1-16. Berlin: Springer.

Piper, A. 2018. Enumerations: Data and Literary Study. 1st ed. Chicago: University of Chicago Press.

Pearce, M. W. 2008. Framing the Days: Place and Narrative in Cartography. Cartography and Geographic Information Science, vol. 35, no. 1, 17–32.

Presner, T. 2016. The Ethics of the Algorithm: Close and Distant Listening to the Shoah Foundation Visual History Archive, In. Eds. C. Fogu, W. Kansteiner, and T. Presner, 175–202. Cambridge: Harvard University Press.

Roux, J. 2009. Telling Lives, Making Place: The Narratives of Three Haitian Refugees in Montreal. MSc diss., Concordia University.

Ryan, M.L. 2012. Space, in The living handbook of narratology (https://www.lhn.uni-hamburg.de/node/55.html) (Accessed Jan. 5, 2021).

Tatman, R. 2017. Gender and Dialect Bias in YouTube's Automatic Captions. Proceedings of the First Workshop on Ethics in Natural Language Processing, Association for Computational Linguistics 2017: 53–59 http://www.ethicsinnlp.org/workshop/pdf/EthNLP06.pdf (Accessed March 4, 2020)

Trumpener, K. 2009. Critical response I. Paratext and genre system: A response to Franco Moretti. Critical Inquiry 36(1): 159-171. doi: 10.1086/606126

Wilkens, M. 2013. The Geographic Imagination of Civil War-Era American Fiction. American Literary History 25 (4): 803-840. doi: 10.1093/alh/ajt045

Watts, P. R. (2010) Mapping narratives: the 1992 Los Angeles riots as a case study for narrative-based geovisualization. Journal of Cultural Geography, 27:2, 203-227. doi: 10.1080/08873631.2010.494401