DETECTING LOCATION NAMES IN FRENCH LIFE-STORY INTERVIEW TRANSCRIPTS

Nadia Bilal

A thesis
in
The Department
of
Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of Master of Computer Science
Concordia University
Montréal, Québec, Canada

September 2021

Concordia University
School of Graduate Studies

This is to certify that the thesis prepared

By: **Nadia Bilal**

Entitled: **Detecting Location Names in French Life-Story Interview Transcripts**

and submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair

(Dr. Olga Ormandjieva)

_____ Examiner

(Dr. Yuhong Yan)

_____ Thesis Supervisor

(Dr. Sabine Bergler)

Approved by _____

(Dr. Leila Kosseim)

Graduate Program Director

_____

(Dr. Mourad Debbabi)

Dean

Abstract

Detecting Location Names in French Life-Story Interview Transcripts

Nadia Bilal

A number of real-world projects cannot leverage the state-of-the-art techniques due to the unavailability of labelled datasets, lack of models tailored to their specific information extraction needs, or lack of models for their language. In such scenarios, instead of using state-of-the-art techniques, a rule-based syntactic analysis is more feasible for extracting specific entities and their relationships. In a similar information extraction scenario, this thesis uses prepositions to detect location names in the French life-story interview transcripts. When the performance is compared with human annotations (gold standard), the average precision for this basic methodology is 80% and the recall is 83%. Such locations that are identified in the context of prepositional phrases are thereafter extracted from the rest of the text. This extends the basic methodology and leads to a significant increase in recall, however, at the expense of precision. The extended version has a higher recall of 94% with a decreased precision of 70%. An additional step addresses a small set of false positives which increases the precision of the extended version to 76% with the same recall of 94%. In addition to location detection, this thesis presents a simple demonstration of using the grammatical context to further detect other entities of interest, specifically, the interviewee's recollection of the past with respect to people in association with a location. Hence, this thesis demonstrates the utility of the rule-based approach and a grammar based methodology to detect specific entities of interest and their relationships in texts of specific projects.

Acknowledgments

I take this opportunity to express my sincere gratitude to everyone who has contributed to this research directly or indirectly.

First of all, I feel indebted to my supervisor, Sabine Bergler, for her valuable knowledge, feedback, motivation, and continuous support throughout this thesis. She introduced me to the field of Natural Language Processing and Computational Linguistics. Her unwavering support and training throughout these years have helped me in establishing a career as a professional computational linguist.

My heartfelt thanks to interviewees, the Page Rwanda Community, the Department of Oral History & Digital Story Telling, the Geomedia lab, and all other people involved in this interdisciplinary research. I would like to especially thank my friend and colleague Emory Shaw for all the discussions, experimentation ideas, and support during the initial phases of this research.

I cannot thank my family enough for their constant encouragement throughout my years of study. This thesis could not have been completed without the patience and never-ending encouragement from my husband Bilal Nagi. I want to express my profound gratitude to my Ammi Jan, my loving sister Mahwish Afzal and my beloved children Maheen Nagi and Muhammad Ali Nagi for their prayers, love, and encouragement.

A big thank you to all my colleagues in the CLaC lab. I would like to especially thank my dear friend and lab mate Sunanda Bansal for her helpful comments and encouraging feedback for this thesis report. I highly appreciate my labmates - Parsa, MinGyou Sung, and Nadia Sheikh for the technical and fun discussions, review, and feedback of this thesis. I would like to express my gratitude towards my dear friends at Concordia University who continuously encouraged me throughout my research work - Banan Qutrieh, Narjesossadat Tahaei, Fathima Lathiff, and Pavel Khloponin. I would also like to thank you all for always being there for me.

My special thanks to my Graduate Program Advisor Halina Monkiewicz, and Racha Cheikh-Ibrahim from Professional Skills Development of Concordia University as well as my LinkedIn mates Dr. Harry Alexander Zwanenburg, Professor Eric Atwell, and Anena Otii for their positive energy and sound advice along the way. Their guidance and mentorship over the years have helped me grow, both personally and professionally. This thesis is a combined result of my abilities and the support of my advisor, my family, and my peers.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In academic disciplines as well as commercial and government organizations, there is often a need to extract specific units of information from specialized corpora. Let's consider a few real-life case scenarios -

Case 1: Detecting PIIs. - A company has a data breach where compromised documents consist of correspondence data among its employees, customers and suppliers. It needs to identify what PIIs (personally identifiable information) are leaked in this attack. A specific set of personal information includes - person name, date of birth, address, telephone number etc. The law requires the company to notify individuals affected by the breach as soon as possible. This, in consequence, requires the company to analyze its data and extract the compromised information precisely.

Case 2: Analysing clinical notes - Oncology researchers want to analyze clinical notes and pathology reports for very specific features and their values over the period of treatments. This information of patient history and records of treatment is buried in natural language text that has automatic processing challenges. The researchers need to detect several features that include - cancer drugs names and their adverse reactions, metastases[1] of cancer to body parts, whether a treatment resulted in a complete cure of the cancer (also called the pathological complete response (pCR)) or not, etc.

Case 3: Location and associated violence - The researchers in digital humanities want to analyze French interview transcripts of genocide survivors. They want to automatically identify the places mentioned by the survivors in their interviews that were associated with violence such as churches, refugee camps, lake sides, forests, villages etc. Moreover, they want to associate the places with the degree of violence, persons and family members involved etc.

The above cases are examples of information extraction tasks. *Information extraction* finds specific entities, relations among entities and events. An example of a well-known information extraction task is the identification of named entities, commonly known as *Named Entity Recognition (NER).*

[1] Metastatic cancer is cancer that has spread from where it started primarily to other parts of the body.

These information extraction needs make it important to understand what constitutes meaningful information to an investigator for appropriate selection of tools and techniques. In order to address the information extraction needs of above cases using the state-of-the-art statistical methods, there are a few concerns -

Specific information extraction requirement - The state-of-the-art statistical methods for information extraction usually focus mostly on named entities. But in the cases above, we can see there are different types of information extraction needs. Moreover, these are very specific to the task and the domain. Most of the existing off-the-shelf state-of-the-art tools are trained for generic entities. For instance in Case 1, among multiple dates present in text specifically the date of birth needs to be identified. Similarly, in Case 2 the clinical notes mention several body parts for different reason and issues. Identifying body parts as metastatic sites is challenging. Therefore, these specific information extraction needs cannot easily be identified using off-the-shelf state-of-the-art tools.

Lack of a labelled dataset - The state-of-the-art methods are often supervised and require a large amount of labelled data to train from scratch. The manual annotation to create labelled data sets is usually considered expensive (Erdmann et al., 2019). Moreover, as we can see in the cases above, the labelling of data is not always feasible. For example, due to privacy and security reasons in Case 1, the data cannot be shared for mass annotation. Similarly, for Case 2, the labelling requires a very specific background in oncology medicine. Therefore, without an existing off-the-shelf trained model and scarce options for training new models, state-of-the-art methods are not always a viable option.

Lack of resources for a language - Several languages lack sufficient data to build supervised machine learning systems (Ehrmann et al., 2016). The most commonly available datasets for NER are for a limited number of natural languages. Therefore, most of the existing trained publicly available state-of-art models are also for these selected languages (Albared et al., 2019).

In scenarios with any of the above concerns, it is often not feasible to resort to state-of-the-art methods for text processing. The popular state-of-the-art statistical methods are mostly used for classification, topic detection, summarization and language translation. However, for tasks like information extraction and entity relationship extraction, the state-of-the-art statistical methods pose some challenges that we discussed above. In such cases, rule based information extraction is a feasible approach that provides appropriate initial results in these scenarios. These and several other projects especially in the clinical domain have use cases where a rule based approach is significantly relevant to the objectives of those projects.

Chiticariu et al., 2013 emphasized that the rule based approaches are prevalent in both academic and commercial projects for information extraction. The need to explore and extract entities and their specific relations is commonly achieved through the rule based approach as it accommodates very well the challenges posed by specific information extractions tasks (Nenadic et al., 2003, Eftimov et al., 2017, Richter et al., 2017, Marciano et al., 2018, Milanova et al., 2019, Law et al., 2019).

In this thesis I assert the validity of a rule based approach. The 3rd case, i.e. the last scenario, begins with location detection in French interview transcripts. It defines a similar scenario for data extraction applications where large labelled training sets are not available. In 2017, at the time of this task, no pre-trained state-of-the-art statistical methods were readily available in our processing environment for location detection in French. Moreover, the information to be extracted for the task includes detection of violence, its degree and association with location and people. This requires identification of very specific entities and extraction of their relations. Therefore, this thesis

- applies a simple rule based approach to detect locations
- evaluates the results with human labelled data
- demonstrates a simple case of extracting information related to detected locations
- asserts that such a methodology is useful in resource constraint scenarios

Chapter 2

Background

The project - The Living Archives of Rwandan Exiles and Genocide Survivors in Canada[1] aims to build a platform with a suite of tools to interact with the life stories of the survivors of 1994 Rwandan genocide. This platform provides an interactive navigation of 28 publicly accessible video interviews through their transcripts. The interviewees mention multiple location names during the interview that have meaningful association to their memories. To present these life-stories on maps such that it shows any mention of violence or of a polar sentiment associated with these locations along with the interview, the Geomedia Lab manually analyzed a few of the transcripts to identify location name mentions as well as, any violence and positive or negative sentiment associated with them[2]. This manual annotation is a labour intensive and costly process. Therefore, the Computational Linguistics Lab at Concordia University assists the Geo Media Lab in automatically detecting such entities. The first and foremost entity to be detected is the name of locations, which is at the center of presenting the information on a map. In this chapter, I will discuss a few concerns and important aspects of this task and the data that play a role in selecting the approach. I will also discuss the suitability of each approach for this task.

The language is either written or spoken. When the spoken language is written down it is known as *spoken text*. In comparison to written text, spoken texts are much more likely to contain grammatical errors, filler words, repairs and repetitions (Blache et al., 2002, Hadži et al., 2012, Wang et al., 2020). The interview transcripts in this project are an example of spoken text in French language. An excerpt from one of the transcripts is given below -

> X.M. : On était six garçons et ..., non... [en train de réfléchir], on était six filles, quatre garçons, pour le moment on est..., cinq filles et un garçon..., trois des garçons ont été tués pendant le génocide et une fille, ma grande sœur, et j'avais aussi une fille à ce temps-là, elle avait l'âge de six ans.
> M.M.: Et puis tes parents ?
> X.M.: Les parents sont là, on a la chance, ils ont survécu..., parmi mes frères et sœurs qui sont morts, l'aîné avait six enfants, alors... les enfants sont là, ils ont survécu. Le père a

[1] https://livingarchivesvivantes.org/about/
[2] http://geomedialab.org/mapping_life_stories.html

été tué, il travaillait..., à Bralirwa-Kicukiro, son corps n'a pas été retrouvé, on pense qu'il a probablement été tué à Nyanza ou à..., proche de Kicukiro de toute façon..., Nyanza ou Gahanga, quelque chose comme ça! Ma fille, ma grande sœur a été tué avec ma fille ..., elle avait un mari et tout curieusement ses enfants ont été... sauvés; alors c'est nous qui les gardons, ce sont des orphelins, ils vivent à la maison, deux filles et un garçon...; le garçon que je suivais directement..., avait une fille mais curieusement, je ne sais pas, dans les circonstances un peu..., pas claire il a été tué après le génocide parce qu'il était rescapé. On dit qu'il a été... empoisonné, vous connaissez notre culture mais..., c'est dur à prouver. Il avait 28 ans ou quelque chose comme ça, alors..., c'est ça. Alors les filles ..., c'est-à-dire les enfants de mon grand-frère, sont là avec leur mère, parce que leur mère a survécu et puis..., là ils sont comme des adultes.

(Translation)
X.M.: We were six boys and ..., no ... [thinking], we were six girls, four boys, at the moment we are ..., five girls and one boy ..., three of the boys were killed during the genocide and one girl, my big sister, and I also had a daughter at that time, she was six years old.
M.M.: And your parents?
X.M.: The parents are here, we are lucky, they survived..., among my brothers and sisters who died, the eldest had six children, so... the children are here, they survived. The father was killed, he was working..., in Bralirwa-Kicukiro, his body was not found, we think he was probably killed in Nyanza or..., close to Kicukiro anyway ..., Nyanza or Gahanga, something like that! My daughter, my elder sister was killed with my daughter..., she had a husband and curiously her children were... saved; so we are the ones who are looking after them, they are orphans, they live at home, two girls and a boy...; the boy I was following directly..., had a daughter but curiously, I don't know, under circumstances a bit..., not clear he was killed after the genocide because he was a survivor. They say he was... poisoned, you know our culture, but..., it's hard to prove it. He was 28 years old or something, so..., that's it. So the daughters..., those are my big brother's kids, are here with their mother, because their mother survived and then..., they are like adults.

The above excerpt from an interview transcript is an example of the spoken text. In the excerpt a few issues have been highlighted. These issues are -

- **Grammatical Errors** - A text is considered grammatically error free if all its sentences are well-formed according to prescriptive grammar rules of its language. These rules include a well defined hierarchical structure of phrases. However, since transcripts are text form of spontaneous speech, the grammatical errors are common.
- **Speaker Turns** - In an interview, the interviewer and interviewee speak in turns. In life-story transcript texts these speaker turns are identified by their initials such as *X.M.:* and *M.M.:* denote speaker turns in above excerpt.

Transcriber Comments - The transcribers add comments to transcripts to capture the details of non-verbal communication. The examples include - *[éternuement]* (sneeze), *[rires]*(laugh), *[En train de réfléchir]*(thinking), *[Accent anglais]*(English accent), *Essuie ses larmes*(Wipes her tears).

Repetitions and self-repairs - When the speaker restates a phrase it is called repetition. Whereas when the speaker says something and then corrects the information in the next phrase then this phenomenon is called self-repairs. In the above excerpt, the interviewee speaks about her siblings as six boys then corrects herself saying that they were six girls and four boys.

Pauses and Filler Words - The pauses and filler words are frequent in spoken text. Some frequently found filler words in life-story transcripts are - *Hmm*, *Alors*, *Ah oui*, *Ben*, *Donc*, *Ok*. Pauses in speech are added into the text through ellipses as ... and in some transcripts by the string << ... >>.

False Start - The false start happens when a speaker starts saying something but stops and leaves the phrase incomplete and rephrases it. In the above excerpt, the interviewee speaks about the killing of her daughter but then pauses and rephrases saying that her elder sister and her daughter were killed. Since transcribers of life-story interviews transcribe every utterance as is, this issue of false starts is frequent in text.

Therefore, spoken texts, such as the interview transcripts, have certain challenges that play an important role in selecting an appropriate methodology for a particular task on such a dataset. To sum it up, these are the important aspects to be noted regarding the dataset -

- French - The transcripts are in French language.
- Spoken text - The transcripts contain grammatical errors, filler words, repairs, repetitions etc.
- Unlabelled - The entities of interest have not been marked in the transcripts.

The task at hand is to extract the entities from transcripts that the researchers at Geomedia lab are interested in, i.e. *entities of interest*. The task of extracting any information from the unstructured[3] text is called *information extraction*. The first information to be extracted from text is the name and mention of a location. The extraction of any named entity, like location, is referred to as *Named Entity Recognition (NER)*. Named Entity Recognition is a particular type of information extraction task. The extraction of named location mentions is Named Entity Recognition task where the only entity of interest is location. The Named Entity Recognition approaches can be broadly classified into three categories - terminology based, statistical and rule based.

The terminology based approach - This approach is appropriate for Named Entity Recognition in the technical corpora that need to detect mentions of fixed sequences of domain specific terms in the text. However, there are challenges with using a terminology based approach for location detection, especially in these Life Stories Interviews. In the beginning of this project, a terminology based approach was used to detect locations by using a GeoNames inspired database that contained

[3]Not in a particular structure like XML etc.

names of locations. This approach fails on homonyms[4]. Too many common French terms were incorrectly marked as locations, when that was not the intended sense of the text. The approach also missed many location mentions as it was realized that in the interview transcripts, the names of the locations mentioned are local, and often not even registered on the map. Moreover, many locations are not referred to by their official names by the locals. Therefore, terminology based approach was deemed too limited and restricted for this task.

The statistical approach - The most recent category of NER approaches use machine learning or artificial neural networks. The best performing models are often supervised, i.e. the features or the weights of features are inferred from the labelled data. In the field of NER, the state of the art are the methods that use supervised artificial neural networks (Yadav et al., 2018; Goyal et al., 2018; Li et al., 2020). However, building a supervised statistical model requires a large amount of manually-labeled training data to generate good results. In such a case, to prepare a labeled dataset, domain experts are needed to perform the task of annotating text for NER labels. For certain datasets and NER tasks, the manual labelling is expensive. Additionally, the datasets are sometimes limited and may not be sufficiently large for building a model which would provide results of good quality. Hence, these NER approaches are usually employed when large public data sets or a large budget for manual labeling is available (Chiticariu et al., 2013).

The machine learning approach is not applicable for named entity recognition and information extraction in interview transcripts for three reasons - First, these models are language-specific and a pre-trained model for French was not readily available in our processing environment. Second, it is also important to note here, that even if a pre-trained model was available, it matters what type of text was used to build the model. The text analysis and information extraction is considered most challenging in spoken text. Since most of the publicly available NER models are trained on written texts, they do not give similar results when tested on spoken language texts. The inherent characteristics and variability of spoken language affects the performance of off-the-shelf tools for text analysis (Chang et al., 2005). Since the model trained on written text may not have come across the disfluencies that we observe in spoken text, the model may not be able to deal with them properly. Therefore, presence of disfluencies like the filler words, repairs and repetitions in the dataset can lead to less-than-optimal predictions. So despite being state-of-the-art, a model built for written texts may produce relatively poor results on spoken texts. Third and most important reason is that location detection is just the first step to mapping the life-stories. But, the goal of mapping a life-story as narrative maps requires not only detection of location name mentions, but also specific other concepts such as violence, positive and negative sentiment and mentions of interviewees relatives using family relationship terms during the interview. Location may be a general Named Entity that the statistical models are often trained to detect. But there are many specific entities that are to be annotated with their relation to a location. For example, which person was where, what violent act happened where, etc. To the best of our knowledge, these concepts cannot be detected with any existing and publicly available ML model for text.

[4]a *"homonym"* is defined as a word that is spelled the same as another word but that does not have the same meaning

The rule-based approaches - The rule-based named entity recognition from text is achieved through rules. In such a rule based approach, the rules are commonly developed by utilizing the context around entities of interest. Most rule based NER systems use dictionaries of names along with rules to capture the context of names. The context helps decide the appropriate category for the name. Besides NER, rule-based systems are usually extended for information extraction such that the next set of rules extract structured relationships among entities of interest from the text. Besides detecting named entities, one can take advantage of the wide array of grammatical constructs to capture relational information among entities. For instance the rule-based NER system, the FUNES (Coates-Stephens, 1992) utilized the grammatical construct in language to detect proper names and their descriptions in text of news articles.

Chiticariu et al., 2013 investigated that although statistical models dominate the research in academia, commercial applications are mostly implemented as rule-based systems. The reason is that to date there are certain NER tasks where named entities and their relations are very specific that require specialized domain knowledge or data that has privacy concerns preventing its distribution. A number of times the datasets are comparatively small but are of significantly high research value to community, government or business. The rule based systems are considered appropriate for such applications. However, though rule based systems can be stand-alone systems, they can also be used in combination with statistical models for high-level feature extraction, solving different pre-processing and post-processing tasks, and semi-automatic creation of gold standards (Kluegl, Atzmueller, et al., 2009; Kluegl, Toepfer, et al., 2016).

However, generally the rule based systems are written with a lot of rules. These rules are usually very specific to the data, text genre, domain and task. Therefore, rule based systems developed for one text genre and showing highly accurate results for it, do not show similar performance on a different text genre. For example, Poibeau et al., 2001 evaluated the performances of two systems developed for the newspaper genre and reported their decreased performance on e-mails and hand-transcribed telephone conversations. Keeping these issues in mind, the methodology of this thesis develops a rule based system with a few, generic and high-level rules for detecting location.

The text corpus of interview transcripts is a typical example of a corpus that consists of valuable information buried in these texts. These texts are required to be processed for extracting location names and other features that are of interest to researchers at Geomedia Lab. However, there exists no labeled data, the corpus size is small and the detection of named entities is exploratory that requires keeping domain experts in the loop. Though it might seem that because unstructured spoken text such as life-story transcript is so varied in form and content, as compared with the written text, that it may not be possible to find grammatical constructions that frequently and reliably indicate the entities of interest. However, in any text, whether spoken or written, grammatical structure of language is always present and words are sequenced in specific ways to deliver the intended meaning. Truly ungrammatical constructions form only a small fraction of the spoken text. The grammatical structure in language can indicate the meanings of lexical items in many useful ways. As established by linguistic literature and research (Coates-Stephens, 1992), the information extraction can rely on syntactic constructs in language and grammar can play a role to extract relevant information units.

It is expected that certain grammatical patterns in these transcripts are consistently present across texts which can be utilized for the purpose of extracting the information of interest. Therefore, this thesis explores a grammar based approach for named entity recognition. In the next chapter I present a systematic methodology based on grammatical constructs to suggest location name candidates in interview transcripts.

Chapter 3

The CLaC Life Stories Pipeline

The entity of interest in the corpus of French interview transcripts, is *location*. To geographers and historians, locations mentioned in transcripts of hour-long interviews are data points of interest. On provision of structured information units containing location names, their narrative mapping tool, Atlascine[1], plots maps of life-story narratives in interviews allowing the end-user to enter the stories using a map interface (Caquard et al., 2014) and make the video interviews navigable. It is possible that the life-story interview transcript contains unofficial and vernacular names for locations, which might not be present in regular location name gazetteers. In order to assist Geomedia Lab in detecting the location names, we developed a Natural Language Processing (NLP) pipeline that analyzes the French transcripts. This pipeline is hereafter referred to as CLaC Life Stories Pipeline (CLSP). The CLSP detects mentions of location names in three phases. The first phase prepares text for syntactic analysis, the second phase does the partial syntactic analysis, and the third phase suggests semantic interpretation of syntactic units. The CLSP pipeline is developed using the General Architecture for Text Engineering (GATE) Framework (Cunningham et al., 2014). An overall brief introduction of GATE is given in Appendix A for reference. Though the methodology in this chapter is discussed independent of the GATE framework, a few GATE modules used in this pipeline are discussed in their respective sections. The upcoming sections discuss each of the three phases of CLSP in detail.

3.1 Preparing Text for Syntactic Analysis

The first phase for syntactic analysis of text is the preparation of text. The preparation involves splitting text into sentences, splitting of sentences into words, removing any unnecessary characters and assigning part-of-speech to words (Palmer, 2000). Besides this, for each of these steps extra considerations are added depending on text and task at hand. The details are discussed in sections that follow.

[1] http://geomedialab.org/atlascine.html

3.1.1 Sentence Splitting

When preparing text for text analysis, sentence splitting is an initial step that identifies sentence boundaries in text. A sentence in any natural language is a unit consisting of one or more words that are grammatically linked to express a statement, question, exclamation, request, command or suggestion. In a natural language processing pipeline, appropriate sentence boundary detection is essential as it affects the efficiency of next modules. Incorrect sentence splitting gives rise to incorrect grammatical class assignment to words by the part-of-speech tagging module and wrong phrase chunking by syntactic phrase chunker.

The CLaC Life Stories Pipeline uses a prepackaged language-independent sentence splitter from GATE (Cunningham et al., 2014) called RegEx Sentence Splitter. It is a regular expression based sentence splitter that detects the sentence boundaries in text by utilizing the punctuation marks and a list of abbreviations that helps in distinguishing end-of-sentence marking periods from the rest. It splits text and identifies sentences on text.

3.1.2 Tokenization

Tokenization is the process of segmenting the characters in text into basic units called tokens which are processed by the subsequent modules in the pipeline. GATE's Unicode Tokenizer (Cunningham et al., 2014) is used to split the text into tokens. It identifies words, numbers, special characters, and punctuation marks. It distinguishes between words in uppercase and lowercase, and between certain types of punctuation marks, thereby conserving essential surface level lexical information that is utilized by next modules. An error in tokenization affects the output of the POS tagging module that assigns grammatical categories to word tokens, and an error in part-of-speech tags affects later syntactic analysis. Since the errors in the tokenization stage have this snowballing effect on the performance of next modules in the pipeline, the tokenizer output is improved to accommodate the spoken text of the interview transcripts and to basic norms of written French Language. A post-processing rule next in the pipeline, handles these concerns as discussed in section below.

Apostrophe Attachment

In written French, the apostrophe mostly replaces one of the vowels when the next word begins with a vowel or a mute h. This apostrophe is attached to the preceding letter. For example the pronoun *Je* is changed to *j'* when next word starts a vowel, the same is true for determiner *Le* that changes to *l'* when next word starts a vowel. Consider the following phrase:

J'ai occupé le poste de Directeur de l'information et des publications.

For the sentence above, the GATE Unicode Tokenizer used in the CLSP generates three tokens for *j'ai* - *j*, *'*, and *ai* - and three tokens for *l'information* - *l*, *'* and *information*; thereby treating apostrophe as a separate punctuation mark. The POS tagger assigns common noun tag to both *j* and *l* when apostrophe is a separate token, whereas *j* and *l* are playing the role of *pronoun* and *determiner* respectively.

This happens because, the part-of-speech taggers are sensitive to tokenization of data in training dataset. If tokenization of text is different from the tokenization of training dataset, the tagger cannot assign correct tags because it does not match it against the tokens it is trained on. For example in case of this apostrophe issue it seems that the French model of Stanford's POS Tagger is trained on dataset where *j'* and *l'* formed single tokens respectively as the detachment affects the decision of part-of-speech tagger. The GATE Unicode Tokenizer is used as-is. However, a simple rule is written to adjust the tokenizer output such that it attached the apostrophe to its preceding string forming a single token *j'* and *l'*, and consequently the tags assigned by the POS tagger changed to *pronoun* and *determiner* respectively.

This pattern is not limited to *j'* and *l'*. It is further corrected by combining apostrophe token with its preceding token whenever a token preceding the apostrophe is one of these: *c*, *d*, *j*, *l*, *m*, *n*, *s*, *t* and *qu*.

Speaker Turns and Transcriber Comments

In life-story interview transcripts, there can be two or more people speaking. These speakers are identified by their initials in text such that their names are replaced by these initials to indicate their turns in the interview transcript. For example, the strings *S.G.:* and *O.G.:* indicate the turns by the interviewer and interviewee respectively in one of the transcripts in life-story corpora. Besides the initials for speaker turns the transcripts also contain certain comments added by the interview transcribers, for example the transcriber adds string in square brackets `[ ]` to add comments and in some transcripts add the string << ... >> into the text to indicate ellipses in speech.

The speaker turns indicated by their initials and the comments added by transcribers are not text. Moreover, these strings break the sentence structure and confuse the part-of-speech tagger which in turn affects the subsequent constituent parsing. Thus it is essential that the initials for speaker turns and comments by transcribers be not processed as text. Hence, as a part of pre-processing these strings for turns and transcriber comments are identified and removed[2] from the set of tokens so as to avoid processing them.

3.1.3 Part-of-Speech Tagging

Once the text has been analyzed into sentences and tokens, each token is then assigned a part-of-speech tag indicating whether the word is a noun, verb, adjective or a preposition etc. A part-of-speech tagger is a module that determines the grammatical category of each word in a natural language text. Although, the state of art for part of speech tagging in text has reached accuracy level that is near human inter-annotator agreement (Manning, 2011), it is also true that most part-of-speech taggers are developed using comparatively well formed texts such as those from news articles, web blogs and Wikipedia articles and therefore are expected to have a lower accuracy on spoken text like interview transcripts. The performance of pos-tagger is better when the sequence of tokens in the text is logical and grammatical. The CLSP uses Stanford POS Tagger (Toutanova et al., 2003) with its off-the-shelf French model available as a GATE plugin, to assign parts of speech tags to

[2]The pre-processing step was developed by Nadia Sheikh, Computational Linguistics Lab at Concordia University

tokens in the text. Table A.1 in Appendix A gives part-of-speech tags by Stanford POS Tagger for French language. Although majority of tokens are assigned correct pos-tags, the errors are possibly due to three reasons:

1. **Consistency in tokenization -** As discussed in the previous section, the POS Tagger is sensitive to tokenization. The tokenizer's output is accordingly adjusted by correcting observed issues. Adjusted tokenization for two issues mentioned in the previous section give consistently improved pos-tags on respective tokens.

2. **Interview Transcripts vs Well Formed Written Text -** The data set consisting of interview transcripts is noisy when compared to the training and test sets of the Stanford POS Tagger (Toutanova et al., 2003). This spoken language text has disfluencies due to repetition, ellipses, and filler words that break the flow of otherwise fluent speech. Therefore, a tagger trained to predict on a written language text may not do as well on a spoken language text, like transcripts.

3. **Unknown Names -** Sometimes the tagger fails to find context that helps it to predict proper noun tag for a capitalized word. Such tokens that are proper names in the life-story interview transcripts get assigned the `ET` tag that is reserved for unknown or foreign token by the POS Tagger.

The issues that were relevant to location detection in transcripts are addressed. Overall, the tags assigned by tagger are not modified, and are used as is for further syntactic analysis. The French model of Stanford POS Tagger is based on Crabbé et al., 2008 commonly called CC tagset with few tag names modified. The parts of speech and their corresponding tags from the tagset that are relevant to the analysis relating to location detection are discussed below -

Noun A word which denotes the name of a person, place, thing or concept in a sentence is called a noun. A noun that names a particular person, place or thing is called a proper noun. Stanford's POS Tagger assigns `NPP` tag to proper noun token in French text. Whereas, a noun that is a generic name for a person, place or thing is called a common noun, and is represented with `NC` tag in tagset. Other nouns that belong to neither the common noun group nor the proper noun group such as *life*, *justice*, *hatred* and *violence* are called abstract nouns and they are tagged `N` by the tagger.

Verb A word denoting an action, occurrence, or state of being is called a verb. `V` tag denotes an auxiliary(also called helping) verb. If the basic form of the verb is detected, it is assigned the `VINF` tag in the sentence. If the past participle form is detected, it is assigned `VPP` tag by tagger. There are other tags for verbs in the tagset however, only these three are relevant for the discussion in this thesis.

Adverb An adverb is a word that commonly limits or restricts the meaning of a verb, an adjective or another adverb. An adverb is assigned `ADV` tag by the tagger.

Adjective A word that typically describes a noun such that it denotes a quality or a characteristic of people, things and phenomena, is called an adjective. It is denoted with `ADJ` tag.

Determiner A word that occurs together with a noun to express the reference of that noun in the context is called determiner. For example, it may indicate whether the noun is referring to a definite or indefinite element of a class, to a particular number or quantity, to a closer or more distant element, or to an element belonging to a specified person or thing, etc. The POS Tagger assigns `DET` tag to the determiner.

Preposition The Stanford POS Tagger assigns `P` tag to a preposition in a French sentence. They generally express a relation in time or space. They can also express relations of agency, cause, means, manner etc.

Pronouns A word that is used in place of one or more nouns is called a pronoun. It is tagged as `PRO`. If specific pronouns are detected then tagger assigns them respective tags. Few pronoun tags used in syntactic analysis rules in next section are: `CLS` for subject clitic pronoun, `PROREL` for relative pronoun, `PROWH` for interrogative pronouns. There are more pronoun tags in this tagset, but only the tags relevant for the discussion in this thesis are mentioned above.

Unknown word When parts-of-speech tagger does not find an appropriate category tag to a word, it marks it as an unknown or foreign word, assigning it `ET` tag in text.

Prefix String The Stanford's POS Tagger identifies certain tokens as prefix strings and assigns them a tag called `PREF`. It is not a part-of-speech, but the tagger assigns it to a prefix token. For example, *"Saint-"* in *"Saint-Catherine"* is assigned the `PREF` tag.

An example demonstrating the application of above tags is provided below. For example, consider the sentence below -

> Á Tripoli, j'ai eu beaucoup de difficultés avec Rudoli, le directeur de l'institution.
> *In Tripoli, I had a lot of difficulties with Rudoli, the director of the institution.*

The above sentence with Part of Speech tags is -

A P *Tripoli* NPP , PUNC *j'* CLS *ai* V *eu* VPP *beaucoup* ADV *de* P *difficultés* N
avec P *Rudoli* NPP , PUNC *le* DET *directeur* NC *de* P *l'* DET *institution* NC . PUNC

Since the text analysis pipeline is developed in GATE, a brief look at the visualization of annotation by GATE is shown in the form of a table in Table 3.1. Table 3.1 shows the Token annotation in GATE for the sentence *"J'ai fait mes études en partie au Congo, en Angleterre, et au Canada."* by the Unicode Tokenizer, along with the category feature added by Stanford POS Tagger. Stanford's POS Tagger has assigned a lexical category to each token by adding *category* feature to the existing *Token* annotation. The table shows Token annotation with features `kind`, `length`, `orth` and `string` from Unicoode Tokenizer. `kind` indicates if token is a word, a number or a punctuation

mark, `length` indicates the length of token, `orth` indicates if the word is in lowercase of starts with an uppercase initial letter, `string` indicates the original string of token as in text. The `category` feature added by POS Tagger indicates the part-of-speech tag assigned to each token. Besides this, for each annotation GATE indicates the start and end off-sets of an annotation within the text which is indicated in `Start` and `End` column in the table. GATE also generates a unique annotation identifier of the annotation which is indicated in the `Id` column of the table. These sequences of tags in a sentence can now be grouped into syntactic phrases through phrase structure rules. With the assumption that the tagger's output is reliable and that each word bears the correct tag, the tags assigned by tagger are used as is for further syntactic analysis.

Table 3.1: Token Annotation from GATE Unicode Tokenizer for sentence: *J'ai fait mes études en partie au Congo, en Angleterre, et au Canada.*

Type	Start	End	Id	Features
Token	1884	1886	1167627	{category=CLS, kind=word, length=2, orth=artapos, string=J'}
Token	1886	1888	1082046	{category=V, kind=word, length=2, orth=lowercase, string=ai}
Token	1889	1893	1082048	{category=VPP, kind=word, length=4, orth=lowercase, string=fait}
Token	1894	1897	1082050	{category=DET, kind=word, length=3, orth=lowercase, string=mes}
Token	1898	1904	1082052	{category=NC, kind=word, length=6, orth=lowercase, string=études}
Token	1905	1907	1082054	{category=P, kind=word, length=2, orth=lowercase, string=en}
Token	1908	1914	1082056	{category=NC, kind=word, length=6, orth=lowercase, string=partie}
Token	1915	1917	1082058	{category=P, kind=word, length=2, orth=lowercase, string=au}
Token	1918	1923	1082060	{category=NPP, kind=word, length=5, orth=upperInitial, string=Congo}
Token	1923	1924	1082061	{category=PUNC, kind=punctuation, length=1, string=,}
Token	1925	1927	1082063	{category=P, kind=word, length=2, orth=lowercase, string=en}
Token	1928	1938	1082065	{category=NPP, kind=word, length=10, orth=upperInitial, string=Angleterre}
Token	1938	1939	1082066	{category=PUNC, kind=punctuation, length=1, string=,}
Token	1940	1942	1082068	{category=CC, kind=word, length=2, orth=lowercase, string=et}
Token	1943	1945	1082070	{category=P, kind=word, length=2, orth=lowercase, string=au}
Token	1946	1952	1082072	{category=NPP, kind=word, length=6, orth=upperInitial, string=Canada}
Token	1952	1953	1082073	{category=PUNC, kind=punctuation, length=1, string=.}

3.2 Shallow Syntactic Chunking

When analyzing a natural language text, we realize that sentences in a language are not just linear strings of words, instead they have an internal structure. Words form groups that combine to form larger groups. These groups of words are called syntactic constituents or syntactic phrases in linguistic literature. These phrases are referenced in the rules of syntactic analysis of any language. To develop a text analysis system that suggests named entities based on grammatical cues, we need to chunk the text into syntactic units. These syntactic units can be detected using a constituency parser. However, the parser provides a full syntactic parse of each sentence that gives a detailed hierarchical relationship among phrases. These details are not helpful when the objective of text analysis is restricted to named entity recognition. Therefore, we select to analyze the text through partial parsing. The partial parsing detects syntactic units such as noun phrases, verb phrases and prepositional phrases, but leaves the decision to connect them for a later stage in text analysis (Allen, 1995). These phrases can then be connected or used by themselves depending on the information extraction needs of the research project. Besides this, partial parsing is considered sufficient for

tasks like named entity recognition (Abney, 1997). Therefore, I decided to write rules for partial text analysis and analyze the spoken text of interviews through a rudimentary syntactic chunker. The next module in the CLaC Life Stories Pipeline (CLSP), identifies contiguous tokens as syntactic chunks based on rudimentary grammar rules determining phrase boundaries. I developed an in-house rule-based chunker purely based on part-of-speech tags and labels of base phrases. This rudimentary chunker gives better control over text chunking and is, therefore, more suitable to the specific text analysis goal of interview transcripts.

3.2.1 Proper Name Chunks

A proper name refers to a specific real world entity. The proper names may consist of a single word such as *Rwanda* or a sequence of words such as *Sainte-Catherine* or *États Unis*. Any such name, must be identified as a chunk and labelled as proper name (`ProperName`) so that it is treated as a single unit in subsequent syntactic analysis. The proper names appeared in French transcripts in following patterns:

1. A contiguous sequence of words carrying proper noun `NPP` tags is a proper name. For example consider the sentence below:

Je	*m'*	*appelle*	*Oscar*	*Gasana*	.
`CLS`	`CLR`	`V`	`NPP`	`NPP`	`PUNC`
			`ProperName`		

   In the sentence above, the two tokens *Oscar* and *Gasana* get `NPP` tags. The sequence forms a proper name *Oscar Gasana* and can be identified as `ProperName`. Examples that include the single word and multiword proper names from interview texts based on this pattern are: *Rwanda*, *Canada*, and *George Henri Levesque*.

2. A sequence of capitalized words with `PREF` tags, by Stanford POS Tagger, when precede an `NPP` tag, or a capitalized word with `NC`, `N`, `VINF` or `ADJ` tag, forms a proper name. This name usually includes one or more hyphens. Let us consider the sentence below:

l'	*université*	*de*	*Dar-*	*Es-*	*Salam*	.
`DET`	`NC`	`P`	`PREF`	`PREF`	`N`	`PUNC`
			`ProperName`			

   In above sentence two tokens with `PREF` tag precede a token with `N` tag, these three tokens together form the name *Dar-Es-Salam*. Examples of some proper names when above pattern is annotated are:*Page-Rwanda* where pos-tag sequence has `PREF` tag for *Page-* and `NPP` tag for *Rwanda*, also in name *États-Unis*, first token carries `PREF` tag for *États-* and `ADJ` tag for *Unis*. Other names that follow this pattern include names: *Trois-Rivières*, *Latour-de-Carol* and *St Paul*.

3. A capitalized word sequence with foreign word `ET` tags is mostly a proper name. Following sentence illustrates the POS tags sequence for multiword proper name - *Mwami Mutara Rudahigwa*

Mwami	*Mutara*	*Rudahigwa*	*a*	*aboli*	*le*	*servage*	.
ET	ET	ET	V	VPP	DET	N	PUNC
ProperName							

   Similarly, other words such as *Umurwanashyaka* and *Nyakivara* are identified as foreign word by the POS Tagger and are assigned `ET` tags, although these are the names of a newspaper and a refugee camp respectively. Hence a sequence of words where each word starts with a capitalized letter and has `ET` tag, is marked as `ProperName`. This way *Umurwanashyaka* and *Nyakivara* are identified as `ProperName`in the respective transcript.

Once this method finds all single and multi-word token sequences in the data that are predicted to be proper names, each sequence is treated as a single unit by subsequent grammars in the pipeline and the text corpus of French interview transcripts is now ready for shallow syntactic analysis. The following sections describe shallow syntactic chunking grammar to identify few relevant syntactic units in the transcripts - noun phrases and prepositional phrases.

3.2.2 Base Noun Phrases

As discussed in section 3.1.3, nouns are words that identify people, places, things, feelings or ideas. The unique names for these nouns are called proper names that we detected in the last section. Although we have nouns and proper names detected in text, there could be words around them that provide further details. These words together form a complete grammatical unit called noun phrase (`NP`). A noun phrase is a syntactic chunk that is centered on a noun. It may contain more than one noun where only one noun is marked as its *head*. Besides a head noun, the noun phrase may have adjectives, adverbs and determiners. The *head noun* is the main concept whereas all other words provide details about the head noun. We detect the boundaries of noun chunks in text and mark the head noun explicitly. Detecting the head played an essential role in location name detection that is discussed in later sections.

Moreover, a noun phrase may embed another noun phrase. In the CLaC Life Stories Pipeline (CLSP), a noun phrase that does not contain another noun phrase is marked as a simple or base noun phrase, whereas one that embeds one or more noun phrases is considered a complex noun phrase. A complex noun phrase that allows maximal length of embedding is marked as maximal length noun phrase. In CLSP, I identified both base and maximal length noun phrases. In the next section, I define rules that identify base noun phrases. The base noun phrases are identified by four rules given below:

1. The first pattern detects pronouns that are identified based on their part of speech tags - `CLS`, `PRO`, `PROREL` and `PROWH` - in text and marks these tokens as base noun phrases. For example

consider the sentence below:

on	*n'*	*a*	*pas*	*de*	*place*	*pour*	*eux*
CLS	ADV	V	ADV	P	NC	P	PRO
BaseNP_Pronoun							BaseNP_Pronoun

In the above example, each token that is assigned `CLS` and `PRO` tags by the pos-tagger is identified as a base noun phrase. Other examples of tokens in text include - *J'*, *ils*, *Elle*, *Tu* and *Nous*. The base noun phrases identified with this rule are marked as `BaseNP_Pro`.

2. The second rule detects bare nouns, i.e. nouns without determiners, as base noun phrases. However, this rule allows optional pre-nominal and post-nominal modifiers. In this rule, a noun with part-of-speech tag `N` or `NC` is marked as the head of the base noun phrase. If there is a token with `ADJ` tag before or after the head noun, it is marked as pre-modifier or post-modifier accordingly.

Maman	*avait*	*une*	*caisse*	*de*	*vêtements*	*différents*
N	V	DET	NC	P	NC	ADJ
BaseNP_BareNoun					BaseNP_BareNoun	

In the above example, *Maman* and *vêtements différents* are identified as base noun phrases. In *vêtements différents*, *vêtements* is marked as the head and *différents* is the post-modifier of this head noun. Other base noun phrases in the text identified with this pattern and, thereafter, marked as `BaseNP_BareNoun` are - *vie*, *problèmes*, *responsable*, *plusieurs personnes* and *Canadien*.

3. The third rule identifies noun phrases with determiners and optional pre and post nominal modifiers. The rule identifies the token with the part-of-speech tag `NC` or `N` as the head noun. Moreover, it covers a broad pattern of pre-modifiers of the head noun. According to this rule, if more than one noun is in a sequence, the last one is marked as the head and those preceding the head noun are considered the pre-modifiers of the head noun. A sequence of multiple adjectives and adverbs preceding the head noun are also considered pre-modifiers. On the other hand, only an adjective or a proper name that succeeds the head noun is marked as its post-modifier. Let us consider the following two examples:

(a)

C'	*était*	*un*	*grand*	*propriétaire*	*foncier*	.
CLS	V	DET	ADJ	NC	ADJ	PUNC
		BaseNP_DetNoun				

(b)

Je	*suis*	*allée*	*à*	*l'*	*île*	*Maurice*	.
CLS	V	VPP	P	DET	NC	NPP	PUNC
						ProperName	
				BaseNP_DetNoun			

In the first example above, *un grand propriétaire foncier* is a base noun phrase and is marked as `BaseNP_DetNoun`. Here, *un* is the determiner, *propriétaire* is the head noun, *grand* and *foncier* are the pre-modifier and the post-modifier of the head noun respectively. In the second example *l' île Maurice* is identified as a base noun phrase and is marked as `BaseNP_DetNoun` as shown above. In this example, *l'* is the determiner and the head *île* is post modified by the proper name *Maurice*. Besides these other base noun phrases that are detected in text based on this rule are - *la Commune Ruhengeri*, *la rue Saint André*, *mon grand-père paternel* and *ma petite sœur Anne-Lise*.

4. The last rule in the base noun phrase chunker detects proper names, which have been identified based on the rules defined in the previous section, as noun phrases. If a determiner precedes the proper name then, as a part of this rule, the proper name along with the determiner is marked as the `BaseNP_ProperName`. In the noun phrases identified using this rule, the `ProperName` is always the head of the noun phrase. For example -

l'	*arrivée*	*massive*	*des*	*réfugiés*	*vers*	*le*	*Rwanda*
DET	NC	ADJ	P	NC	P	DET	NPP
							ProperName
						BaseNP_ProperName	

In the above example, *le Rwanda* is a base noun phrase, marked as `BaseNP_ProperName`. As you can see from the example above, all the word sequences that were marked as `ProperName` in previous section are marked with `BaseNP_ProperName` such that the preceding determiner is also part of new span.

To sum up the rules defined so far, the base noun phrases (`BaseNP`) in text are identified in four possible ways expressed as following rewrite rule:

$$\texttt{BaseNP} \rightarrow \texttt{BaseNP\_Pronoun} \mid \texttt{BaseNP\_BareNoun} \mid \texttt{BaseNP\_DetNoun} \mid \texttt{BaseNP\_ProperName}$$

3.2.3 Base Prepositional Phrases

A prepositional phrase (`PP`) consists of a preposition and a noun phrase. The preposition is the head of the prepositional phrase and the noun phrase that follows the preposition is called the complement

of the preposition. Once base noun phrases (BaseNP) have been identified, the identification of prepositional phrases is straightforward. Similar to the base noun phrase, a base prepositional phrase is one that does not embed another prepositional phrase in it. A simple rule marks the prepositional phrase as BasePP in the text where a token with the part-of-speech tag P precedes a base noun phrase (BaseNP).

$$\texttt{BasePP} \rightarrow \texttt{P BaseNP}$$

A few examples of prepositional phrase identified in text based on above rule are given below -

(a) *à* *l'* *île* *Maurice*

P DET NC NPP

BaseNP

BasePP

(b) *avec* *ma* *femme*

P DET NC

BaseNP

BasePP

(c) *au* *Québec*

P NPP

BaseNP

BasePP

3.2.4 Maximal Length Noun Phrases

An important property of natural languages is that they allow *"recursion"*, that is a sentence, considered the longest syntactic constituent, consisting of constituents that themselves may consist of smaller constituents. Let us consider the constituent parse tree of a noun phrase as given below -

We can see in Figure 3.1 parse tree that the noun phrase constituent recursively contains other noun phrases constituents, showing recursion within syntactic constituents. The syntactic chunker developed for this methodology accommodates this recursive process allowing a noun phrase and a prepositional to contain one or more prepositional phrases recursively. Since shallow syntactic chunking loses the hierarchical structure that is present in a parse tree, the maximal length noun phrase has a flat structure instead, as in Figure 3.2.

So far, we have identified base noun phrases and base prepositional phrases in text. In the next step, I identify the maximal length noun phrase (MaxNP) as the longest noun phrase that consists of

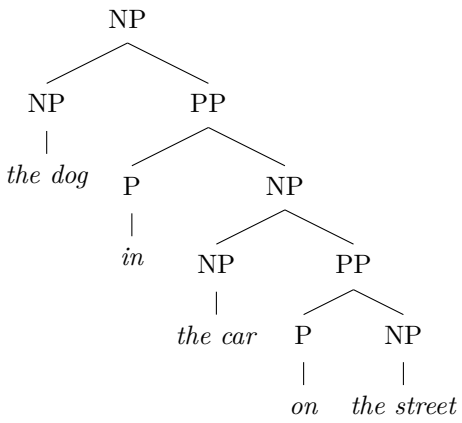


Figure 3.1: Parse tree of a phrase demonstrating recursion in noun phrases

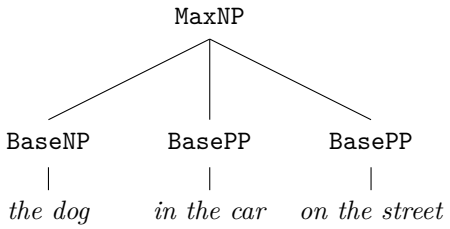
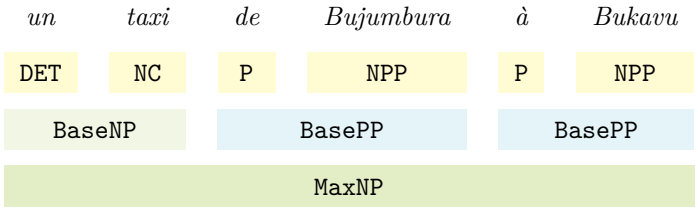


Figure 3.2: Flat structure of a maximal length noun phrase (`MaxNP`) detected through partial parsing

a base noun phrase followed by base prepositional phrases. Following rule in our shallow syntactic grammar detects maximal length noun phrases.

$$\texttt{MaxNP} \rightarrow \texttt{BaseNP BasePP+}$$

Table 3.2 presents some maximal length noun phrases as identified by the above rule in text. The head of the maximal length noun phrase is the head of the base noun phrase that precedes the modifying prepositional phrases. Consider the maximal noun phrase - *un taxi de Bujumbura à Bukavu*, here the head of the phrase is *taxi*, where *de Bujumbura* and *à Bukavu* are the modifying prepositional phrases. Identifying the head of noun phrases supports their semantic interpretation in the next section. Following illustrates the marking of this maximal length noun phrase in text.



la région de Gitarama
un passeport de réfugiée Burundaise
Le directeur de l'école
le pouvoir à Kampala
Le systéme politique en Tanzanie
l'université de Dar-Es-Salam
votre départ de l'Éthiopie vers la Libye
la Commission Économique des Nations Unies pour l'Afrique
les maisons sur les collines en face de Bunyambiriri
un taxi de Bujumbura à Bukavu
la communauté des rescapés du génocide
Tunisie pendant le période de guerre

Table 3.2: Example maximal length noun phrases from text

3.2.5 Maximal Length Prepositional Phrases

A commonly occurring phrasal construct in life-story narratives is of two or more prepositional phrases in a sequence. These prepositional phrases provide context for locative or temporal aspects of events which can be useful in next levels of analysis. Similar to recursion explained using parse tree for a noun phrase constituent in figure 3.1 and its corresponding flat structure 3.2 achieved through shallow syntactic chunking, figure 3.3 and figure 3.4 show recursion in prepositional phrases -

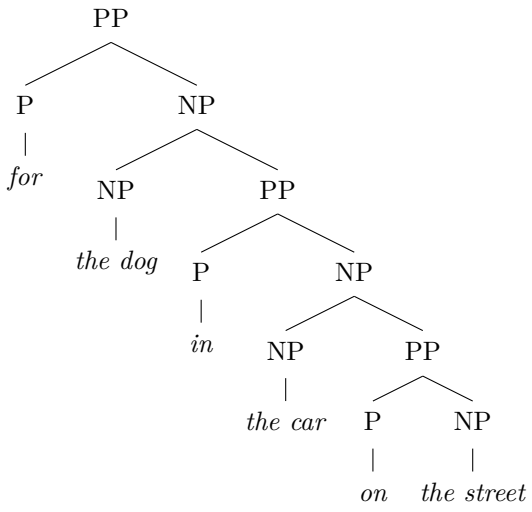


Figure 3.3: Parse tree of a phrase demonstrating recursion in prepositional phrases.

Whenever a preposition precedes a `MaxNP` the text span is marked as maximal length prepositional phrase (`MaxPP`). Mark the text chunk where the part-of-speech tag of the token preceding a maximal length noun phrase is `P` as `MaxPP`. The rule to mark maximal length prepositional phrases is as

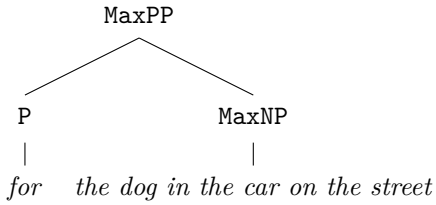
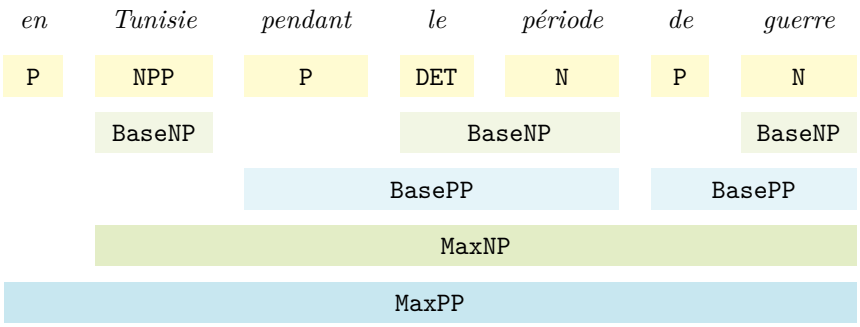


Figure 3.4: Flat structure of a maximal length prepositional phrase (`MaxPP`) detected through partial parsing.

follows -

$$\texttt{MaxPP} \rightarrow \texttt{P MaxNP}$$

Table 3.3 presents some prepositional phrases of maximum length from transcripts. Following demonstrates the detection of maximal length prepositional phrases in text -



The grammar presented in this section for syntactic analysis utilized parts-of-speech tags and token orthogonality features to detect syntactic phrases. The next section builds on top of these structures, such that the syntactic context is made specific in order to drive specific interpretation of textual units and thereby extracting information contained in text.

à Kinshasa avec ma femme
à l'Université du Québec à Trois-Rivières
À l'université du Québec à Montréal
au Nations Unies à Tripoli dans l'église pendant des semaines
dans la communauté des rescapés du génocide au Rwanda pour la première fois en 1985
dans ce mémorial de Murambi
après les vacances de Noël en 1961
à l'Université du Québec à Trois-Rivières
à Gikongoro dans la commune de Karama sur une colline
au Québec en 1980 avec ma famille

Table 3.3: Example maximal length prepositional phrases from text

3.3 Semantic Interpretation of Syntactic Units to Identify Locations

There are many ways that the structure of a language can indicate the meanings of lexical items. But the difficulty lies in finding constructs that consistently and reliably indicate the relation of interest. It is perceived to be easier to find such regular structures in well formed written text than in spoken text. The unrestricted text such as interview transcripts is not only noisy but also so varied in form and content, that finding such constructs in them seem difficult if not impossible. Fortunately, every text contains some naturally occurring structural elements called syntactic phrases, such that identifying them help find such regularities. In the previous section, we have identified syntactic chunks in text.

This section presents how meaning is assigned to these chunks occurring in specific positions in the text.

The syntactic phrases encode grammatical relations among words and provide a guideline to understand the text. For instance, a group of words chunked as a noun phrase indicates that this chunk is mentioning some place, things or concept. The position of this noun phrase within a clause determines its function as subject, object or oblique object in clause. This way shallow syntactic analysis can guide towards shallow semantic interpretation. This section analyzes text to prescribe some meaning to grammatical chunks such that the interrelationships among syntactic constructs is examined and their semantic interpretations are demonstrated.

The main theme of the life-story narrative in a transcript is the forced displacement and compelled relocation of interviewees and their families during the 1994 genocide. As described in the previous chapter, the detection of mentions of location names in these narrative texts is of primary importance. The text spans where location names are mentioned can be detected through grammatical signals in text. Some prepositions are a strong indicator of presence of location names (Vandeloise, 1991). Therefore, the next step is to create a list of these prepositions that are considered locative in nature. In the next section, I present a few such prepositions that are used in identification of locations in this methodology.

3.3.1 Locative Prepositions

In French text locations are often indicated by certain prepositions, `à` and `de` for example. The head of their complement noun phrase can be interpreted as a location. Moreover, though certain other prepositions do not predominantly indicate a location, they do have location names as their object sometimes (Table 3.4). Below I discuss two prepositions and their lexical variants -

à/au/aux - The preposition **à** takes the form **au** and **aux** depending on the gender and count of its complement noun. Vandeloise, 1991 gives the prescriptive use of the French preposition **à** that it almost always designates its complement proper name as a location name. In following examples we demonstrate that these prepositions indicate location names -

1. *je suis rentrée avec ma tante* *à* `P` *Kigali* `LOC`

à	au	aux	avant
d'	de	du	dans
en	entre	par	près
sous	sur	vers	via

Table 3.4: Prepositions for location detection

2. *Au* P *Rwanda* LOC *il était architecte*
3. *vous allez aux* P *États-Unis* LOC

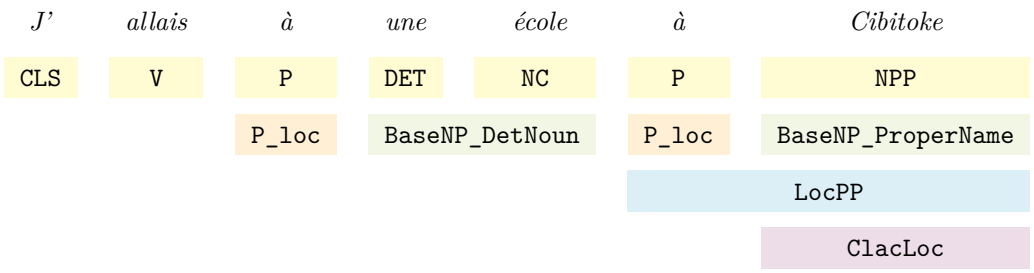
d'/de/du - Th preposition *de* marks attributive relationships between two noun phrases, where one or both noun phrases can be location names or can be names of other entities. However, unlike *à*, it can not be considered as an explicit location indicator (Carlier et al., 2013). But this preposition has frequently taken a location name as its object in transcripts. Therefore, the preposition *de* and its variants are used for identifying locations in this methodology. Below are some examples of sentences with prepositions *de*.

1. *Il vient d'* P *Italie* LOC
2. *je leur ai acheté une propriété prés de* P *Nyanza* LOC
3. *vous avez le plus de souvenir du* P *Brundi* LOC

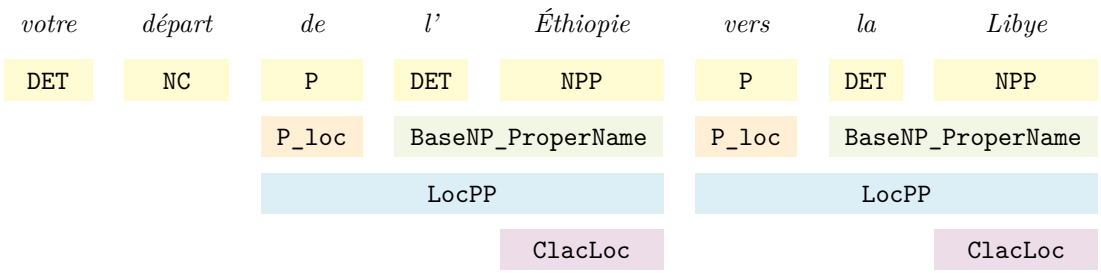
Besides `à` and `de`, other prepositions that indicate location names in life-story transcripts are given in Table 3.4. Among all the prepositions in the French language, some prepositions are explicitly excluded from this list such as `avec` and `chez`. These prepositions are excluded as they are often indicative of person names instead of location names. In the next section we use the prepositions listed in Table 3.4 to detect location names in interview transcripts.

Location Detection

The Table 3.4, presents prepositions that are collected as a list and are considered locative prepositions `P_loc`. Based on this, a simple rule identifies those base prepositional phrases where the head preposition is one of the locative prepositions and the complement noun phrase has a proper name as its head. When we identified base noun phrases in the previous section the rule (`BaseNP_ProperName`) is the only one that marks a proper name as its head. These base prepositional phrases are marked as locative prepositional phrases `LocPP` and the proper name is predicted as location name and marked as CLaC Location (`ClacLoc`). Below I demonstrate the detection of location names using prepositional phrases -



In the above example the preposition*à* appears more than once in the sentence. In the first phrase, it has the noun phrase `une école`, the `BaseNP_DetNoun`, as its complement. The head of this compliment phrase is a common noun *école*. As described earlier there are two restrictions for location name detection. First, a preposition, `P_loc`, from Table 3.4 must be there. And second the head of the complement noun phrase should have been identified as a proper name. This phrase matches the first condition but fails for the second one. This way *école* is not marked as a `ClacLoc` location. On the other hand, *Cibitoke* is head of noun phrase detected with rule `BaseNP_ProperName`, therefore *Cibitoke* is detected as a location name and marked as `ClacLoc`. Next



The example above has two locative prepositional phrases. In the first prepositional phrase, the complement noun phrase of preposition is *l'Éthiopie*. Similarly in the second prepositional phrase the complement noun phrase is *la Libye*. As I described in the previous section, only the head of the complement noun phrase is identified as the location name. Therefore, the determiners in these noun phrases do not become part of the location name and only *Éthiopie* and *Libye* are detected as location names and marked as `ClacLoc` respectively.

As I mentioned in last section that the preposition *à* can be considered a reliable indicator of location names in text but we observe in above example that other prepositions such as *de* and *vers* also have location names as complements. The Table 3.5 show more sample phrases from transcripts that demonstrate this phenomenon of prepositions, though not explicitly considered location indicators, consistently indicate location mentions.

However, using this methodology, though the proper names that appear in the prepositional phrase are identified as location, the same name outside the prepositional phrase context is not detected as a location. In the next section, I will address this issue by a simple technique.

elle a traversé *sur* ***Bujumbura*** LocPP *avec le bébé*
Il est passé *par* ***Kampala*** LocPP
De ***Bukavu*** LocPP *tout le long du lac, jusqu'* *à* ***Goma*** LocPP
À ***Tripoli*** LocPP *j'ai eu beaucoup de difficultés avec Rudoli le directeur de l'Institution*
Ils disaient tous les gens *de* ***Gikongoro*** LocPP
En ***Belgique*** LocPP *c'était beaucoup plus problématique*
C'est tout près *de* ***Nyanza*** LocPP *vers* ***Butare*** LocPP *au sud* *du* ***Rwanda*** LocPP
Je n'ai pas une idéée de la distance *entre* ***Saint-Pierre*** LocPP *et l'autre côté.*

Table 3.5: Sample locative prepositional phrases identified in the texts. The bold text inside a locative prepositional phrase (`LocPP`) is a location name.

Extended Location Detection

In the last section, we identified the proper names that are direct complement of locative prepositions as `ClacLoc` location in text. However, the same location name is not detected when it appears outside this syntactic context. Therefore, in order to extract these occurrences, the proper names detected as `ClacLoc` location are collected into a list. Afterwards, this list is used to extract the occurrences of these proper names in text as the mentions of a location. These occurrences are marked as Extended ClacLoc (`ExtCLoc`). Below I demonstrate this step of extending the location name detection outside prepositional phrases -

1. *je ne pouvais pas passer* *à* ***Orangegugu*** LocPP
2. *Non ma demi-sœurs travaille* *à* ***Gisenyi*** LocPP
3. *De* ***Bukavu*** LocPP *tout le long du lac, jusqu'* *à* ***Goma*** LocPP
4. *Ça joint Goma* et *Bukavu, côté Congolais et Gisenyi et Orangegugu, côté Rwandais.*

In example sentences (1), (2) and (3) above, proper names *Goma*, *Bukavu*, *Gisenyi* and *Orangegugu* are detected as location names as they appeared in prepositional phrases where head of the respective phrase is one of the prepositions in the Table 3.4. In Example 4, these occurrences are not identified as locations since they are outside the syntactic context. Therefore, to detect these occurrences that lie outside the context of prepositional phrases, the `ClacLoc` location names are collected in a list. This list is then used to detect all instances of location names in the respective transcript and mark them as `ExtCLoc` as shown below -

1. *je ne pouvais pas passer à* *Orangegugu* ExtCLoc
2. *Non ma demi-sœurs travaille à* *Gisenyi* ExtCLoc
3. *De* *Bukavu* ExtCLoc *tout le long du lac, jusqu' à* *Goma* ExtCLoc
4. *Ça joint* *Goma* ExtCLoc et *Bukavu* ExtCLoc, *côté Congolais et* *Gisenyi* ExtCLoc *et* *Orangegugu* ExtCLoc, *côté Rwandais.*

Besides extending the location name detection beyond the syntactic context, this step accommodates another important aspect that pertains to partial names mentioned in text. Usually, a location name that consists of more than one token can be mentioned in text in its full or partial form. If part of the name was identified as a `ClacLoc` location, the full name of the same location should also be detected in text as location. For example, if *Addis* is detected as `CLacLoc`, then its full name *Addis-Abeba* should also be identified as the location name. This is achieved by identifying the entire `ProperName` as a location as long as it fully or partially contains a name listed in the list of `ClacLoc` locations. The following example demonstrates the partial name detection and its extension to full location name detection -

1. *jeunes réfugiés qui étaient* ***à Addis*** `LocPP`
2. *j'ai quitté* *Addis* `ExtCLoc` *pour Tripoli*
3. *parce qu'* *Addis-Abeba* `ExtCLoc` *c'était quand même un milieu politisé*

In Example 1, *Addis* is detected as `ClacLoc` and collected as a part of the list that extends the location detection. As we can see in Example 2 and Example 3, when this list is used to detect *Addis* for its mentions in the transcript, all of its mentions with and without prepositional context are detected as location names `ExtCLoc` including the full name *Addis-Abeba*.

Furthermore, a number of times the names detected as `ClacLoc` locations are not, in fact, location names. Therefore, if these names, that are incorrectly identified as the names of locations, are collected for extending location name detection, it will only further propagate the error. The following example presents a case where a name is incorrectly detected as a `ClacLoc` location -

1. *il a été commis par les Hutu*

   P DET NPP
   P_loc BaseNP_ProperName
   LocPP
   ClacLoc

2. *c' était un hutu*

   DET NC
   BaseNP_DetNoun

3. *des réfugiés hutu*

   DET NC ADJ
   BaseNP_DetNoun

In the Example 1, following the process of `ClacLoc` location identification, *Hutu* is identified as a location since *Hutu* is the head of the noun phrase that complements the preposition *"par"*. Although the word *Hutu* is identified as a proper name in Example 1, the same word appears in text

as a common noun (Example 2) and as an adjective (Example 3). In the Example 2, *hutu* describes one individual of the class of *Hutus*. Furthermore, as an adjective in Example 3, it qualifies or characterizes the noun *réfugiés* as *Hutu*.

As we can see, identifying *Hutu* as a location name is already incorrect and extracting its other occurrences in text as location mentions will further propagate the error. In order to avoid this propagation of error, a restriction is imposed that location names will only be extended to `ProperName` phrases that contain a name from the list of `ClacLoc` locations. Therefore, in the examples above, while the occurrence of *Hutu* in Example 1 is still erroneously extracted as a location mention, it is not extracted in Example 2 and Example 3.

In our dataset, once the `ClacLoc` location name list has been inferred from each transcript, all the proper names in that transcript that exactly or partially contain a name from the list, are extracted as the mentions of location names. Hence, at the end of this step the `ExtCLoc` marks all mentions of location names in a transcript based on its respective list of `ClacLoc` location names. In the upcoming chapters we will not only evaluate this location detection methodology, but also demonstrate a simple case of extracting information related to the locations detected in this chapter.

Chapter 4

Comparative Evaluation and Results

In the last chapter, we have discussed the approach to detect locations in the French interview transcripts. In this chapter, we will evaluate the results of this methodology. The upcoming sections present the set up of evaluation, the results and their discussion.

4.1 Evaluation Setup

Before we discuss the results, we must discuss the set up of experiments for the evaluation and the metrics that are used for a comparison of performance.

4.1.1 Gold Standard

In order to evaluate the results of CLaC Life Stories Pipeline (CLSP) discussed in the previous chapter, we need a Gold Standard. A native French speaker who is not otherwise associated with the project has manually annotated 16 transcripts for mentions of the location names where a location name is name of any concrete place mentioned in the transcript. The manual annotator was required to label all occurrences of names of locations in each transcript. This labeled set of transcripts forms the Gold Standard Corpus for evaluation of CLSP location detection.

4.1.2 Evaluation Measures

When comparing the locations predicted by CLSP against the gold standard, following three outcome cases are possible -

1. **True Positive (TP) case** - For a location detected by CLSP there is a corresponding location label in gold standard.
2. **False Negative (FN) case** - For a location detected by gold standard there is no corresponding location name identified by CLSP. That is the CLSP missed the location name.

3. **False Positive (FP) case** - For a location marked by CLSP there is no corresponding location label in the gold standard.

On the basis of these three cases, we can measure the performance of locations using CLSP with the classical entity-level evaluation metrics: precision (P), recall (R) and f-measure (F). The precision measures the fraction of correctly identified mentions out of total identified mentions. Since, the correct mentions identified are true positives (TP), and the total mentions identified are sum of true positive (TP) and false positive (FP), the precision can be defined as -

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4.1}$$

The recall is the fraction of correctly identified mentions out of total correct mentions. As for precision, the correctly identified mentions are true positive (TP). However, the total number of correct mentions is sum of true positive (TP) and false negative (FN). Therefore, the recall can be defined as -

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4.2}$$

The f-measure, also called balanced F-score (F1 score), is the harmonic mean of precision (P) and recall (R) -

$$\mathrm{F1} = 2 \times \frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}} \tag{4.3}$$

Among these evaluation measures precision is preferred for systems where the cost of wrong prediction is high. On the other hand, the recall is preferred in systems when cost of missing is high. The higher the recall rate, the fewer the missed entities. The CLSP prefers recall over the precision as it is easier for Geographers to discard a wrongly detected location than to locate a missed mention of location.

When evaluating location detection, the mention identified by the CLSP and that mentioned in the gold standard may -

1. completely overlap, or
2. partially overlap, or
3. not overlap at all

While calculating the performance scores, each location mentioned contributes full weight for complete overlap and half weight for partial overlap. The complete overlap of identified locations with gold standard is called a strict match, whereas, the partial overlap is called a lenient match (Maynard et al., 2006). In this chapter, I will evaluate the CLSP for both strict and lenient matches.

4.2 Evaluation Results

This section presents the results of the methodology executed by CLaC Life Stories Pipeline (CLSP). I analyzed the results to give insights into both the particularities of the data as well as the potential and limitations of the CLSP for location detection. I report both strict scoring (exact overlap of results obtained from CLSP and gold standard annotations) and lenient scoring (partial overlap of results obtained from CLSP and gold standard annotations).

4.2.1 Evaluation of the Second Set of Manual Labels

Another human annotator was arranged to manually label the same set of transcripts. The second set of manual labelling was done on fifteen out of sixteen transcripts where the transcript `RI-MU` was not labeled. Two native French speakers have annotated the transcript corpus for location mentions. As a first comparison, I have compared this second set of labels with the gold standard labels. This comparison is presented in the 4.1.

	Location	
TRANSCRIPT ID	strict	lenient
AN-MU	95	95
EM-KA	93	94
BE-HA	90	91
ST-GA	84	87
MA-JO	86	86
L-C3	84	88
OS-GA	94	96
RI-MU	-	-
EM-HA	88	89
YV-IS	89	89
LI-ND	92	93
MA-CH	90	92
FR-C3	87	87
XA-MU	80	83
BE-KA	86	93
PH-MU	84	85
average	88	90

Table 4.1: Comparison of locations identified by the second annotator with the Gold Standard

Table 4.1 shows that the strict f-measure by second annotator is in the range of 80-95 whereas the lenient f-measure is between 83-96. The fact that the f-measure is at least 80 percent and goes up to 95 percent shows that there is a high degree of overlap between the location mentions labelled by the two annotators. However, even though location might seem a simple concept to

understand, the difference in the understanding of two annotators reflected in labelling results show that the definition of location may vary among different annotators. This shows that even for people, description of location may vary. This analysis is helpful in putting the location detection results of CLSP in context.

4.2.2 Baseline - ANNIE

Besides evaluation against the gold standard, I compared the results of the location detection by CLSP against a baseline. The baseline in my experiments is the French version of Named Entity Recognition pipeline, called ANNIE (A Nearly New Information Extraction system), which is available as default named entity recognition pipeline with the General Architecture of Text Engineering (GATE) framework. The ANNIE pipeline consists of 5 modules - sentence splitter, tokenizer, POS Tagger, set of lists (gazetteers) and a set of rules defined in GATE's rule description language[1]. These modules, in their order of execution in the pipeline, are -

GATE Unicode Tokenizer - It tokenizes the document at white space and punctuation marks.

RegEX Sentence Splitter - The second module splits text into sentences.

Stanford's POS Tagger - The French version of ANNIE uses Stanford's POS Tagger with its French language model. It assigns part-of-speech tags to each token.

ANNIE Gazetteer The ANNIE gazetteer is a collection of lists that include - tiles, person names, organization names location names and list of abbreviations used with organization and location names etc

Named Entity Tagger NE transducer consist of rules to identify multiple named entities that include - Person, Location, Organization, Date, Money and Percent annotations

Orthomatcher provides orthographic coreference that is it matches proper names and their variants in a document.

There are also other modules available in the ANNIE plugin, which are not used in the default application, but can be added if necessary ANNIE identifies multiple entities that include - Person, Location, Organization, Date, Money and Percent annotations. For the purpose of this evaluation we are interested in only Location annotation by ANNIE as our baseline.

Table 4.2 presents the evaluation results of our baseline (ANNIE) against the gold standard for lenient match and strict match for precision, recall and f-measure. In the next section, these results are used for comparative analysis of baseline and results of CLSP.

4.2.3 ClacLoc vs Baseline

As discussed in the previous chapter the location names are identified in the transcripts based on the syntactic context. These location mentions are marked as `ClacLoc`. Table 4.3 compares `ClacLoc` locations detected by CLSP to the locations identified by the baseline (ANNIE).

[1] GATE documentation for ANNIE

Text ID	Baseline					
	strict			lenient		
	P	**R**	**F**	**P**	**R**	**F**
AN-MU	84	74	78	90	79	84
EM-KA	76	58	66	84	64	72
BE-HA	70	69	70	86	85	86
ST-GA	79	78	78	87	83	85
MA-JO	83	62	71	86	64	73
L-C3	75	73	74	79	77	78
OS-GA	82	79	80	92	88	90
RI-MU	78	72	75	84	78	81
EM-HA	74	73	74	83	82	82
YV-IS	63	62	62	69	68	69
LI-ND	72	66	69	81	74	78
MA-CH	78	60	68	86	67	75
FR-C3	72	71	71	80	78	79
XA-MU	62	58	60	71	65	68
BE-KA	79	64	71	81	66	73
PH-MU	78	68	73	88	77	82
average	75	68	71	83	75	78

Table 4.2: Evaluation Results of Baseline (ANNIE)

Text ID	strict						lenient					
	Baseline			ClacLoc			Baseline			ClacLoc		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
AN-MU	84	74	78	94	87	90	90	79	84	94	87	90
EM-KA	76	58	66	88	81	85	84	64	72	89	81	85
BE-HA	70	69	70	92	68	78	86	85	86	93	69	79
ST-GA	79	78	78	84	82	83	87	83	85	85	83	84
MA-JO	83	62	71	82	85	84	86	64	73	82	85	84
L-C3	75	73	74	86	86	86	79	77	78	89	89	89
OS-GA	82	79	80	85	82	84	92	88	90	85	82	84
RI-MU	78	72	75	81	87	84	84	78	81	81	87	84
EM-HA	74	73	74	75	86	80	83	82	82	75	86	80
YV-IS	63	62	62	71	86	78	69	68	69	71	86	78
LI-ND	72	66	69	87	74	80	81	74	78	87	74	80
MA-CH	78	60	68	81	87	84	86	67	75	81	87	84
FR-C3	72	71	71	73	85	79	80	78	79	73	85	79
XA-MU	62	58	60	70	84	77	71	65	68	71	85	77
BE-KA	79	64	71	70	90	79	81	66	73	70	90	79
PH-MU	78	68	73	68	81	74	88	77	82	69	83	76
average	75	68	71	80	83	82	83	75	78	81	84	82

Table 4.3: Comparison of Baseline (ANNIE) vs ClacLoc

As we can see in Table 4.3, the average score of ClacLoc is better for both lenient and strict match for each evaluation metric except for precision in lenient matching. The performance of baseline is very low for all three measures for the transcripts `YV-IS` and `XA-MU`. For `YV-IS` and `XA-MU` transcripts, the recall of the baseline is 58 and 62 with strict matching, respectively. However, the recall for `YV-IS` and `XA-MU` using `ClacLoc` with strict matching is increased to 84 and 86, respectively. This is shows that `ClacLoc` leads to at least 35% improvement[2] in strict match and at least 25% improvement over the baseline with lenient matching. The average improvement of `ClacLoc` over the baseline in f-score is around 15% for strict match and 5% for lenient match. Another important aspect to note is that average f-measure score for baseline (ANNIE) improves nearly 10% when matched leniently. Whereas the average f-measure of `ClacLoc` remains exactly the same for both strict and lenient matching. The improvement in score when matching leniently suggests that a number of times location mentions detected by baseline are partially overlapped with gold standard whereas no difference in the case of `ClacLoc` suggests that the locations identified by `ClacLoc` mostly identifies the full location name in the location mention.

As we discussed in the previous chapter, in the methodology used to identify `ClacLoc` locations, the location name detection is limited to the syntactic context of prepositional phrases. To extract

[2]Percentage of improvement is calculated as $\Delta\% = \frac{\text{New Value}-\text{Old Value}}{\text{Old Value}} \times 100$

Text ID	Baseline			ClacLoc			ExtCLoc.		
	P	R	F	P	R	F	P	R	F
AN-MU	84	74	78	94	87	90	93	95	94
EM-KA	76	58	66	88	81	85	87	95	91
BE-HA	70	69	70	92	68	78	86	95	90
ST-GA	79	78	78	84	82	83	78	93	85
MA-JO	83	62	71	82	85	84	76	96	85
L-C3	75	73	74	86	86	86	78	91	84
OS-GA	82	79	80	85	82	84	72	96	83
RI-MU	78	72	75	81	87	84	72	96	83
EM-HA	74	73	74	75	86	80	72	96	82
YV-IS	63	62	62	71	86	78	67	95	78
LI-ND	72	66	69	87	74	80	73	85	78
MA-CH	78	60	68	81	87	84	61	96	75
FR-C3	72	71	71	73	85	79	59	93	72
XA-MU	62	58	60	70	84	77	57	97	72
BE-KA	79	64	71	70	90	79	48	98	64
PH-MU	78	68	73	68	81	74	47	92	63
average	75	68	71	80	83	82	70	94	80

Table 4.4: Strict comparison among Baseline (ANNIE), ClacLoc, and Extended ClacLoc

mentions of location names irrespective of their syntactic context, this methodology is extended to detect mentions of `ClacLoc` locations in the entire document. In the next section I evaluate the performance of the results obtained with this extension and compare its results with location mentions identified by the Baseline and `ClacLoc`.

4.2.4 ExtCLoc vs ClacLoc vs Baseline

As described in the Methodology chapter, in CLSP the location names are first identified using locative prepositional phrases as `ClacLoc` locations. A list of these location names is then used to capture their occurrences in respective transcripts irrespective of syntactic context as `ExtCLoc`. Table 4.4 and Table 4.5 presents comparisons of `Baseline`, `ClacLoc` and `ExtCLoc` for strict and lenient matches, respectively.

The Table 4.4 shows `ExtCLoc` shows a significant improvement of 38.2% over the baseline in terms of average recall. However, the average precision of `ClacLoc` is 80 which drops more than 10% for `ExtCLoc` locations. It means extending the location detection beyond syntactic context has significantly improved recall by compromising on precision. But as we discussed in the beginning of this chapter, a higher recall is preferred for location detection in this task. Therefore, the performance of `ExtCLoc` is in line with the objective of our task. Moreover, though the precision has decreased with the extension of `ClacLoc`, it can be improved by increasing the precision of location names

Text ID	**Baseline**			**ClacLoc**			**ExtCLoc.**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
AN-MU	90	79	84	94	87	90	94	96	95
EM-KA	84	64	72	89	81	85	88	96	92
BE-HA	86	85	86	93	69	79	88	97	92
ST-GA	87	83	85	85	83	84	80	95	87
MA-JO	86	64	73	82	85	84	78	98	87
L-C3	79	77	78	89	89	89	83	97	89
OS-GA	92	88	90	85	82	84	73	97	83
RI-MU	84	78	81	81	87	84	73	97	83
EM-HA	83	82	82	75	86	80	72	97	83
YV-IS	69	68	69	71	86	78	68	97	80
LI-ND	81	74	78	87	74	80	74	86	80
MA-CH	86	67	75	81	87	84	62	96	75
FR-C3	80	78	79	73	85	79	60	94	73
XA-MU	71	65	68	71	85	77	58	99	73
BE-KA	81	66	73	70	90	79	48	99	65
PH-MU	88	77	82	69	83	76	49	96	65
average	83	75	78	81	84	82	72	96	81

Table 4.5: Lenient comparison among Baseline (ANNIE), ClacLoc, and Extended ClacLoc

identified as ClacLoc. As we will see in the next section, a simple filtering of ClacLoc names before extending them as ExtCLoc will improve the precision of location detection. As we observed in Table 4.3, that the average f-measure between lenient and strict score of ClacLoc remained same, similarly there is less than 2 percent change in average f-measure for ExtCLoc in strict and lenient scores.

4.2.5 Extended ClacLoc (Excluding Black List) vs Extended ClacLoc vs ClacLoc

The gazetteer and exclusion lists are the most intuitive way for researchers to tailor their NLP environment. In order to improve the precision of location mentions detected with Extended ClacLoc (ExtCLoc), I observed the list of names extracted using ClacLoc and found some names in the list that are clearly not location names. A list of a total of five such strings, referring to the people of Rwanda, account for many false positives in ExtClacLoc. These strings in the transcripts are: Rwandaise, Tutsi, Hutu and their plurals Tutsis and Hutus respectively. These strings formed an explicit exclusion list which is called the *Black List*). These names were removed from the ClacLoc list of location names before extracting their occurrences in the transcripts. The location mentions identified through the process of filtering out Black List names before extending were marked ExtCLoc+. Table 4.6, evaluates the performance of locations identified with ExtCLoc+ against the methodologies discussed so far.

Table 4.6 shows a comparison of the results of ExtCLoc+ with the baseline. As expected, Table 4.6 when compared with Table 4.4 for ExtCLoc show that the average precision has increased by at least 5% in both strict and lenient matching. Moreover, in order to compare with the labels of the second annotator, let's look at the average f-measure of location mentions detected in the 15 transcripts (except RI_MU). The average f-measure of ExtCLoc+ for these 15 transcripts is 83.86 and 84.93 for strict and lenient matching respectively, whereas the average f-measure of the second annotator's location labels is 88 and 90. These numbers show that ExtCLoc+ is clearly outperformed by the location mentions labelled by the second annotator, but by a margin of only 6% at most.

4.2.6 Overall Performance

Table 4.7 presents the overall performance of the baseline along with ClacLoc, ExtCLoc and ExtCLoc+. This table shows the average values for strict and lenient matches for each of these methodologies for each evaluation measure. Along with average values, the table also provides a comprehensive view of the variability of these scores throughout the transcripts for each measure. It is important to note that this performance evaluation is for these sixteen transcripts and it might be different on other texts. We can see that the average precision for strict match is maximum for ClacLoc. Also, it has a reasonably high f-measure with tightest variability around the average for this corpus. It is thus the best suited procedure with balanced precision and recall. On the other hand, if recall is preferred, as is the case with this project, then ExtCLoc+ gives maximum recall and offers a reasonable trade-off in precision and the variability of f-measure.

Text ID	strict						lenient					
	Baseline			ExtCLoc+			Baseline			ExtCLoc+		
	P	R	F	P	R	F	P	R	F	P	R	F
AN-MU	84	74	78	94	95	95	90	79	84	95	96	95
BE-HA	70	69	70	88	95	91	86	85	86	90	97	93
EM-KA	76	58	66	87	95	91	84	64	72	88	96	92
L-C3	75	73	74	82	91	86	79	77	78	85	95	89
OS-GA	82	79	80	82	96	88	92	88	90	82	97	89
ST-GA	79	78	78	81	93	87	87	83	85	83	95	89
MA-JO	83	62	71	76	96	85	86	64	73	78	98	87
RI-MU	78	72	75	77	96	86	84	78	81	78	97	86
EM-HA	74	73	74	75	96	84	83	82	82	75	97	85
MA-CH	78	60	68	73	96	83	86	67	75	74	96	83
YV-IS	63	62	62	71	95	81	69	68	69	72	97	82
PH-MU	78	68	73	67	92	78	88	77	82	70	96	81
LI-ND	84	74	78	75	85	79	87	74	80	76	86	81
FR-C3	72	71	71	67	93	78	80	78	79	68	94	79
XA-MU	62	58	60	61	97	75	71	65	68	62	99	76
BE-KA	79	64	71	55	98	71	81	66	73	56	99	71
average	76	68	72	76	94	84	83	75	79	77	96	85

Table 4.6: ExtCLoc+ vs Baseline (ANNIE)

4.3 Qualitative Analysis of Results

In this chapter, so far we've analysed the results of location detection using CLSP in a quantitative fashion. However, we have yet to go beyond the numbers and take a closer look at the results to identify potential areas of improvement of CLSP for location detection. In this section, we will take a closer look at cases where locations were detected incorrectly or missed altogether.

4.3.1 False Positives

As presented in the Methodology chapter, the `ClacLoc` location names are detected in CLSP using the syntactic context of proper names. In a locative prepositional phrase (`LocPP`), the head of base noun phrase, i.e. a `ProperName`, is interpreted as a location name and marked as the `ClacLoc` location. However, some names that are identified as a location, in fact are not referring to a location. These cases are considered false positives for this task. We can see in Table 4.7 that the average precision for strict matching of location mentions identified as `ClacLoc` locations is 80%. This implies that nearly 20% of the proper names marked as `ClacLoc` location, are not in fact mentions of a location name. These 20% cases form the false positives for `ClacLoc` locations. Since the extensions of `ClacLoc`, i.e. `ExtCLoc` and `ExtCLoc+`, are based on the identification of location

	Baseline		ClacLoc		ExtCLoc		ExtCLoc+	
	strict	**lenient**	**strict**	**lenient**	**strict**	**lenient**	**strict**	**lenient**
Precision	76^{+8}_{-14}	83^{+9}_{-14}	80^{+14}_{-12}	81^{+13}_{-12}	70^{+23}_{-23}	72^{+22}_{-24}	76^{+18}_{-21}	77^{+18}_{-21}
Recall	68^{+11}_{-10}	75^{+13}_{-11}	83^{+7}_{-15}	84^{+6}_{-15}	94^{+4}_{-9}	96^{+3}_{-10}	94^{+4}_{-9}	96^{+3}_{-10}
F-Score	72^{+8}_{-12}	79^{+11}_{-11}	82^{+8}_{-8}	82^{+8}_{-6}	80^{+14}_{-17}	81^{+14}_{-16}	84^{+11}_{-13}	85^{+10}_{-14}

Table 4.7: Overall Performance Results for Location Annotations

names from `ClacLoc` locations, I will first discuss the common patterns in false positives observed in `ClacLoc` locations.

Consider the results of transcripts `BE-KA` and `PH-MU` in Table 4.4. As you can see from the table, these two transcripts show a significant drop in precision when `ClacLoc` is extended to `ExtCLoc`. For these transcripts, the precision drops from 70% and 68% for `ClacLoc` to 48% and 47% for `ExtCLoc` respectively. When analyzed closely, the names that contributed the highest percentage to false positives are recorded in Table 4.8 and Table 4.9. These tables present sample phrases with names that are falsely identified as location mentions for transcripts `BE-KA` and transcript `PH-MU`. It is clear from Table 4.8 this that mere three strings account for almost one third of the false positives in this text. Out of these three strings *Félicité* and *Thomas* are the names of people instead of location. *Gacaca*, on the other hand, is the name of the tribal judiciary. Since these names were incorrectly identified as `ClacLoc` locations, their extension `ExtCLoc` further propagated this error leading to drastic drop in precision. Similarly, the four names that have the maximum contribution in the false positives for transcript `PH-MU` are: *Habyarimana*, *Kagame*, *Gacaca* and *Pâque*. Just like the false positives in transcript `BE-KA`, the top contributors in this transcript, i.e. *Habyarimana* and *Kagame*, are instead the names of people. These two names by themselves account for 28% of the false positives for this transcript. It was observed that the names of people account for a large portion of false positives for each transcript. The question is why?

Name	**#**	**Context**
Félicité	15%	…pour ne pas répondre à la question *de* ***Félicité*** `LocPP` …le frère *de* ***Félicité*** `LocPP` …
Thomas	12%	…je prends les mains *de* ***Thomas*** `LocPP` …Omar demande *à* ***Thomas*** `LocPP` …
Gacaca	5%	…je n'ai jamais participéé *au* ***Gacaca*** `LocPP` …

Table 4.8: Top 3 false positive names and their percentage in total number of false positives for BE-KA transcript

Some prepositions strongly indicate locative interpretation of their complement proper name.

Name	#	Context
Habyarimana	16%	...les ficelles sont tirées par la famille *de* ***Habyarimana*** LocPP ...
Gacaca	6%	...le résultat *de* ***Gacaca*** LocPP c'est ça ...
Pâques	5%	...fêter la fête *de* ***Pâques*** LocPP ...
Kagame	4%	...c'est le drapeau *de* ***Kagame*** LocPP ...

Table 4.9: Top 4 false positive names and their percentage in total number of false positives for PH-MU transcript

However, it is observed in Table 4.8 and Table 4.9 that the other names, especially person names, are also appearing as the complement of such prepositions. In my list of locative prepositions (Table 3.4), the most frequently occurring prepositions in the texts are *de, d' (from, of)* and *à (to)*. But this subset of prepositions is shared for both location and person. Furthermore, it was observed that the preposition that frequently contributed to the incorrect detection of `ClacLoc` location is `de`. Though this preposition introduces a location in text, it also frequently marks attributive relationships between two noun phrases. Therefore, though sometimes the complement of the preposition `de` is a location name, it is not exclusive to locations. This pattern is observed in other transcripts as well and thus shows the polysemous nature of prepositions in French. This polysemous nature of prepositions is ultimately the reason for a large proportion of false positives being a name, just not of a location.

Additionally, in some cases, the names identified as `ClacLoc` locations by CLSP were not even proper names, but instead the common names of different entities. For example -

1. Beaucoup *de* ***Rwandais*** LocPP parlent aussi le Swahili ...la culture *Rwandaise* ExtCLoc
2. il y a un bar *pour les* ***Hutu*** LocPP ...Il prenait quelques étudiants *Hutu* ExtCLoc ,
3. il y a des familles *de* ***Tutsi*** LocPP ...Tous les *Tutsi* ExtCLoc vont mourir

The names such as *Tutsi*, *Hutu* and *Rwandaise* are specifically of note because of their contribution to overall false positives. An analysis of `ExtCLoc` results for all transcripts indicated that 5 names - *Tutsi*, *Tutsis*, *Hutu*, *Hutus Rwandaise* - account for an average of nearly 21% of all the false positives. Rather than names of location, these are common names referring to a group of people. The people of Rwanda are collectively called *Rwandaise* and Hutu and Tutsi are the ethnic majority and minority groups among the *Rwandaise* people respectively. However, in the above examples, these strings are considered proper nouns by the part-of-speech tagger. But these names are clearly not proper names. In the CLSP methodology the identification of proper names forms the basis for the location names finally identified. Therefore, though we proceed with the assumption that the part-of-speech tags provided by part-of-speech tagger are accurate, practically, it is seldom true. Therefore, this error can be considered a consequence of part-of-speech tagging error. This error depends entirely on the performance of part-of-speech tagger.

4.3.2 False Negatives

Though some names were incorrectly identified as a location, some location names were missed entirely. These location mentions that are marked in Gold Standard as location names but are missed by CLSP are considered false negatives. A name identified as a location name in Gold Standard can be missed by the `ClacLoc` location detection for one of two reasons - incorrect tagging or lack of its use in a locative prepositional phrase.

As discussed in the previous section, the part-of-speech tags assigned to words may not be entirely correct. If a proper noun is incorrectly assigned a common noun `NC` or abstract noun `N` tag, it is not identified as a proper name. However, in the CLSP methodology the identification of proper names forms the basis for the location names finally identified. Therefore, one reason for when the name sequence cannot be detected as a location is the error on the part of part-of-speech tagger.

The second reason for a missed location mention is that the location name never appeared in any locative prepositional phrase in the respective transcript since CLSP is designed to look for the names of location within the preposition phrases only. For example -

1. *je dois venir* *au Canada* `LocPP` .
   Translation : I have to come to Canada.
2. *j'avais appris dans mon cours de géographie que le Canada c'est…, c'est froid*
   Translation : I learned in my geography class that Canada is …it's cold

Although *Canada* was detected as `ClacLoc` location in the first sentence, it is missed in the second example due to missing syntactic context. If, in the entire text, *Canada* never once occurs in a locative prepositional phrase, then it will not be identified as the name of a `ClacLoc` location and therefore will not be detected as a location mention in the subsequent modules of CLSP. Therefore, such a name will not be identified as `ExtCLoc` location and this location name will be a false negative not only for `ClacLoc` but also for `ExtCLoc`. A few examples of such locations are given below -

1. *dans le quartier où on était …, Muhima près de chez Kabuga …, près du …, de la route poids-lourd qu'on appelle …, ou quelque chose comme ça, il y avait la tension, …*
   Translation : in the neighborhood where we were …, Muhima near Kabuga's place…, near the …, the truck road we call …, or something like that, there was tension,…
2. *c'était le camp de…, ou les camps, parce qu'il y en avait plusieurs mais il y en avait un qui s'appelait je crois Nyakivara.*
   Translation : it was the …camp, or the camps, because there were several of them but there was one called I believe Nyakivara.

In the examples (1) and (2), *Muhima* and *Nyakivara* are names of a neighbourhood and a camp respectively. Since these two location names never occurred in a locative prepositional phrase (`LocPP`), they were not detected as location names by `ClacLoc` or `ExtCLoc`. Similarly, in another transcript *France* and *Allemagne* never occurred in a (`LocPP`) and missed getting detected as location names.

The results and discussion so far show that some prepositions are stronger grammatical cues for mentions of location names than others. With minimal and generic syntactic information the CLaC Life Stories Pipeline (CLSP) has achieved high quality results for this spoken text corpus. These results support our initial hypothesis that detecting the mentions of an entity of interest can be guided by grammatical cues in language. Moreover, Table 4.4 and Table 4.5 present that simple `ClacLoc` achieves highest f-measure for both, strict and lenient scoring respectively. It is thus the procedure best suited for holistic scoring with balanced precision and recall. However, the location detection using CLaC Life Stories Pipeline (CLSP) can also be adjusted in favour of either precision or recall. This can allow for opportunities beyond those that could be achieved if only F1 (which is maximized by equal recall and precision) was targeted.

In this chapter, we have evaluated the locations detected using CLSP. However, as mentioned before, many real world information extraction tasks require related information to be identified. In the next chapter, we will discuss a very simple case of detecting related information to reiterate the value of grammatical analysis in building simple yet effective solutions for complex tasks.

Chapter 5

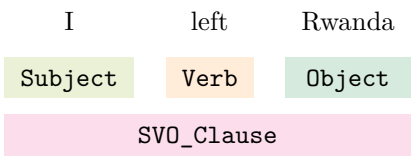
Who and Where - Detecting more than just location

The text of life-story interview transcripts contains a treasure of information. Visualizing the locations in grammatical context opens up numerous opportunities to see them in various aspects. Apart from locations, in the transcripts there is a lot of other information about people, family members, dates, times, movements, sentiments, violence etc. The researchers at Geomedia Lab are interested in extracting this information and their relations to the locations. Since the goal of the project is to create cartographic maps of each interviewee's life-story narrative, it is important to capture the information specific to each interviewee. For example, where were they born, where did they spend their childhood, where did they study, where were they displaced to during the genocide, which locations are related to their experience of violence etc. In this chapter, I will briefly discuss a simple methodology applied to extract only one type of information of interest mentioned before, i.e. mentions of interviewee or their family in relation to a location. Moreover, the mentions of people are only extracted when mentioned in reference to the past.

In the previous chapters, we have discussed a simple methodology to extract locations through syntactic analysis. These mentions of location names can be connected to other entities in a meaningful way through a syntactic chunk called Subject-Verb-Object (SVO) clause, which we will discuss below. Since our focus is on associating people with locations, in this chapter we will only identify the mentions of people in the transcripts that are connected with the mention of location names.

5.1 Subject-Verb-Object (SVO) Clause

A group of phrases, usually centred around a verb phrase is called a *clause*. A clause that describes the relationship between subject and object in terms of verb is called a Subject-Verb-Object (SVO) clause. For example, in the sentence below, the relationship between the subject represented by the pronoun *I* is related to the object *Rwanda* is given by verb *left* -

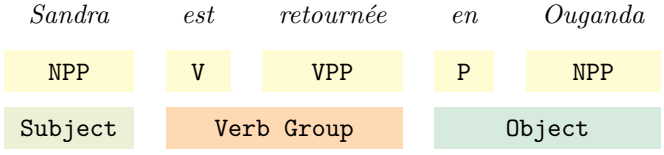


As we can see in the example above, by identifying the SVO clauses in the text, we can capture the relationship of location mentions with other information (or entities) of interest.

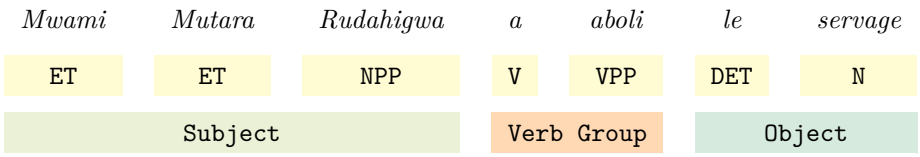
An SVO clause has three components - subject, verb and object. The *subject* of a sentence or clause shows the topic of sentence, or who or what is performing an action in the sentence. The *verb* tells what action is done by the subject. Therefore, in simplistic terms, the entire SVO clause tells that some action (given by the verb) was taken by the *subject* with respect to the *object*, and thereby defines the relationship between the subject and the object in terms of the verb. In an SVO clause, the *subject* and the *object* often refer to the syntactic position of noun phrase in relation to the main verb. Since we have already detected the noun phrases in previous chapters, the next step is to detect the verb in the clause. However, often the action, or the state, or the occurrence given by the verb is further defined by additional words around it. Therefore, for a rudimentary identification of SVO clause for our purposes, we first identify the entire *verb group* that serves the purpose for defining the action. Some of these words are essential for referring to time of the action, state or the occurrence. Such words that are used with the main verb to help express the main verb's tense are called the *auxiliary* or the *helping verbs*. As mentioned before, the information of interest is the one mentioned in reference to the past. For this purpose, we use the grammatical cues of *passé composé* tense in French. The *passé composé* in French describes an action, a specific event or a succession of specific events that occurred in the past. In the text, the auxiliary verbs are assigned tag `V` by the part-of-speech tagger, whereas the words in the past participle form are assigned the tag `VPP`. Therefore, past tense verb groups can be extracted as the sequence of auxiliary verb `V` followed by past participle form of a verb `VPP`. Additionally, the verb group may be optionally preceded or/and succeeded by adverbs (`ADV`). Therefore, a rudimentary grammar to detect the `VerbGroup` indicating past tense in French can be given as -

$$\texttt{VerbGroup} \rightarrow \texttt{ADV? V+ VPP ADV?}$$

For the current information extraction task, the verb group is identified based on the above grammar. A few sentences below demonstrate the use of above grammar for identifying the entire verb group -

1. 

2.

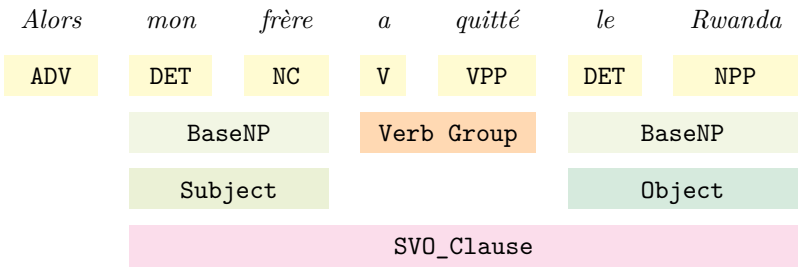


Based on the `VerbGroup` identified above, the SVO clause can be defined as a Noun Phrase followed by a Verb Group which is followed by either a Noun Phrase or a Prepositional Phrase. Therefore, in terms of the base and maximal length noun phrases and prepositional phrases, SVO clause can be defined as in the grammar given below-

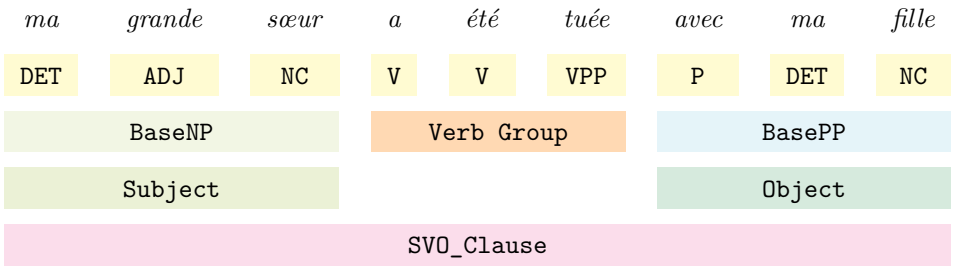
`SVO_Clause` → `(BaseNP | MaxNP) VerbGroup (BaseNP | BasePP | MaxNP | MaxPP)`

A few examples of the identification of an SVO clause based on the above grammar and verb group are given below -

1.



2.



The SVO clauses identified with the above grammar will have only the subject-verb-object relationships that are mentioned in reference to the past. However, the information of interest is limited to specific case where in the SVO clause

- the interviewee or people related to interviewee are mentioned
- at least one location identified as `ExtCLoc` is present
- the utterance refers to the past

In the next section, we'll use the SVO clause identified in this section to extract the information of interest.

5.2 Who and Where

In this chapter, the goal is to identify the mention of people in that are mentioned in connection with a location in the interview transcripts. The SVO clause identified in the section before is generic, which means it gives us all entities in a subject-verb-object relationship mentioned in the reference of past. However, for the current goal, we are interested in the subject-verb-object relationship between people (subject) and locations (object) only. We have already identified location mentions using CLSP in the previous chapters. But we have yet to identify people. In the sections that follow, I'll first identify people and then extract clauses where people and locations are connected together to extract the desired information.

5.2.1 Identifying people mentions

In the transcripts, the people are usually talked about in one of these ways -

Name The transcripts have mentions of people by their names. These includes the names of family members of the interviewee, their friends, the names of people involved in politics or in the genocide etc. For example, a few samples from the text where people are referred to by their names are -

- ***Solange*** *sont nées au Congo*
- ***Sandra*** *habitent ici au Canada avec moi.*
- ***Papa*** *travaillait au Burundi*
- *l'avion du* ***président Habyarimana*** *était abattu*

Family relation In the interviews, the interviewees often talk about their childhood, marriages, problems faced by their relatives during the genocide etc. Therefore, people are often mentioned through their family relations in the interview transcripts as shown below -

- *parce que* ***leur mère*** *a survécu*
- *je ne compte pas aller enseigner comme le fit mes...,* ***mon grand-père*** *paternel,* ***ma grand-maman*** *maternelle,* ***mon père*** *et* ***ma mère*** *et* ***mon frère****.*
- *Et j'étais avec Immaculée,* ***mon épouse*** *actuelle.*

Groups Due to the narration of conflict between the ethnic groups - Tutsis and Hutus - the people are also referred as a group. This includes the mention of people by their nationalities such as Belges, Tutsi, Hutu, Interahamwe, Rwandaise, Canadien etc. A few samples from the transcripts where people are referred to as a group are given below -

- *je ne savais pas s'ils étaient* ***hutus*** *ou* ***tutsis****...*
- *il a été commis par* ***les Hutu***
- ***les congolais*** *n'avaient pas ce complexe d'infériorité..*
- *l'arrivée massive* ***des réfugiés Burundais*** *vers le Rwanda*

Pronoun In the transcripts, the speaker often uses the pronouns to refer to people mentioned earlier in discourse by their names, family relations or tribe name. For example -

- *... parce que quand **J'** avais 20 ans, la plus vieille, **elle** avait cinq ans, six ans...*
- *Donc **ils nous** considéraient comme des criminels,...*
- ***ils** ont massacrés beaucoup de membre de sa famille.*

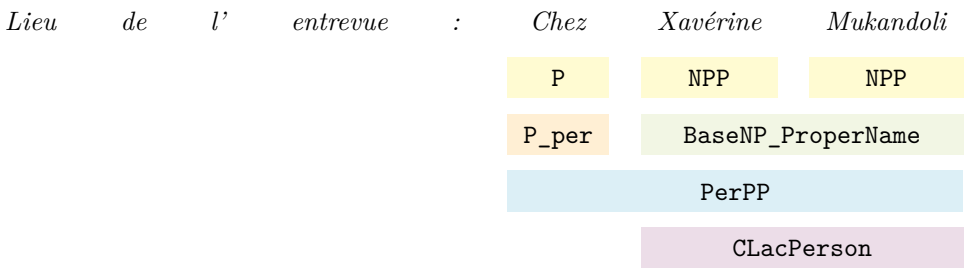
However, we are only interested in the mentions of interviewee themselves or people related to them. So, we'll not identify mentions of people as a group. So, in this section we'll approach the identification of mention of people by their names, pronouns or family relations -

People mentioned by their pronouns - People are referred to by several pronouns like, *il*, *elles*, *je*, *nous* etc. However, the current goal is to identify the mention of interviewee or people related to them only. Moreover, we are interested in the pronouns that can be connected to locations in an SVO clause. Therefore, we are interested in first person pronouns only that can be the subject of an SVO clause, i.e. *je* and *nous*. Therefore, these pronouns were considered mention of interviewee and people related to them and marked `ClacPerson` for the extracting the information we need.

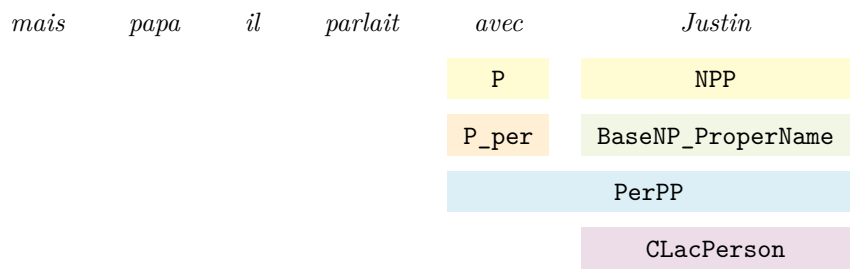
People mentioned by their names - In the first step of CLSP's shallow syntactic analysis, we have already identified the proper names (`ProperName`). A proper name includes most named entities mentioned in an interview transcript, like the names of locations, people, organizations etc. In CLSP, we identified some of the proper names as locations. However, it was observed that most of the proper names are either referring to location or people. Therefore, for simplicity, all the proper names that were not identified as a location mention (`ExtCLoc`) were assumed to be referring to people.

Additionally, in CLSP, the locations were extracted based on the prepositions which were considered locative. However, there are two preposition *chez* and *avec* which were observed to be distinct in the sense that their complement proper name is mostly a person name. In interview transcripts, we interpret the proper names appearing as complements of these two prepositions as candidates for people names and are marked as `CLacPerson`. These two prepositions are marked as `P_per` and respective prepositional phrase is marked as `PerPP`. For example -

chez In the prepositional phrases where the preposition *chez* is the head of the phrase, its complement is often the name of person. For example, in the sentence below, the complement of the preposition *chez* is *Xavérine Mukandoli*, which is the name of a person -



avec The French preposition *avec* is used in similar way as the English preposition *with*. When a proper name appears as its complement, it carries the meaning of accompaniment. Therefore, often the complement of this preposition is the name of a person. For example, in the sentence below, the complement of the preposition *avec* is the name of a person, i.e. *Justin*.



Finally, the proper names identified as people names were marked `CLacPerson`. Table 5.1 shows a few phrases from the transcripts where proper names were identified as the names of people (`CLacPerson`) (given in bold).

le président ***Kayibanda***
le major ***Kanyarengwe***
le juvénat ***Sainte-Trinité***
la commune ***Gishyita***
sa fille ***Mukasonga***
monsieur ***Costier***
l'île ***Ijwi***
le roi ***Mutara Rudahigwa***
s'appelait ***Aphrodis***
ENTREVUE AVEC ***OSCAR GASANA***

Table 5.1: A list of some phrases from the transcripts where the proper names are identified as the names of the people (given in bold).

As we can see in Table 5.1, some of the names do not refer to a person. For example, *Ijwi* in *l'île* ***Ijwi*** is the name of an island, not a person. Similarly, it was observed that some names of people were missed, because a person's name was not identified as a proper noun (`NPP`) by the POS tagger.

beau-fils	belles-sœurs	enfants	frères	mamans	oncles	père
beau-frère	cousin	enfants	garçon	mari	papa	pères
beau-père	cousine	famille	garçons	maris	papas	soeur
beaux-frères	cousines	familles	grand-mère	mère	parent	soeurs
beaux-pères	cousins	femme	grand-mères	mères	parents	sœur
belle-fille	demi-frère	femmes	grand-parent	neveu	petit-enfant	sœurs
belle-mère	demi-frères	fille	grand-père	neveus	petit-fils	tante
belle-sœur	demi-sœur	filles	grand-pères	nièce	petite-fille	tantes
belles-filles	demi-sœurs	fils	grands-parents	nièces	petites-filles	
belles-mères	enfant	frère	maman	oncle	petits-enfants	

Table 5.2: A list of some terms for family relations in French

Moreover, it was observed that two family terms - Papa and Maman, appear in similar context as proper names across multiple transcripts. The capitalized instances of both of these terms are often identified as a proper noun (`NPP`) by the part-of-speech tagger. Therefore, these terms, despite referring to people by their family relations, are marked as proper names.

People mentioned by their family relations - When interviewees narrate their life-stories, the family members are mostly mentioned by their family relation names instead of their personal names, for example, *ma sœur*, *ses parents* etc. In the transcripts, most of the references to the family members, like *sœur* and *parents*, are often preceded by possessive pronouns, such as *ma*, *mon*, *sa*, *son*. This observation was used to identify the mentions of family members. However, since the goal in this chapter is to identify the mention of people related to interviewee only, I extracted the family member mentions based on the first person possessive pronouns (`FPPP`) only, i.e. *ma*, *mon* and *mes*, *nos* and *notre*. Therefore, there are two components for identifying the reference to family members -

Possessive Pronouns The possessive pronoun used to refer to family member - ma, mon, mes, nos, and notre.

Family Terms A list of family relations given in Table 5.2

Based on the components above, the following rules were used to identify mentions of family members -

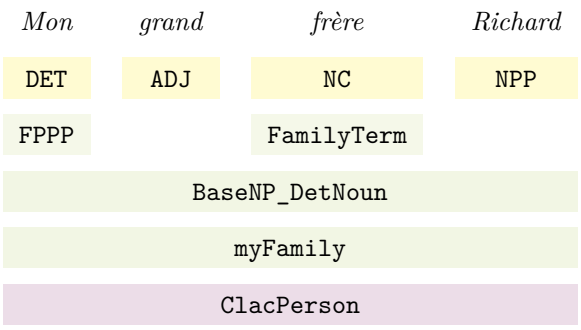
1. **Immediate family** In this rule, the mentions of immediate family relations are identified. The goal here is to identify the Base noun phrases (`BaseNP`s) that contain the mention of immediate family. In a noun phrase, the main concept is given by the *head noun* whereas all other words provide details about the head noun of the noun phrase. It is observed that the family terms are usually the head noun of a base noun phrase. Moreover, for the purpose of identifying the relationship of the interviewee with the family relation mentioned, we also need the base noun phrases with the possessive pronouns, like *mon*, *mes* etc. Therefore, we

Mes parents
Mon papa
mes nièces
mon épouse
nos deux tantes
nos tantes maternelles
notre troisiéme fille
mes autres grands frères
mon frère Jean-Pierre
ma petite sœur Anne-Lise

Table 5.3: A list of some phrases from the transcripts with the mentions of immediate family members of the interviewee.

use the base noun phrases with determiners (`BaseNP_DetNoun`) for extracting a subset of these phrases which mention immediate family relations. Therefore the noun phrases that represent the immediate family members of the interviewee can be identified as the ones that match both of the following conditions -

- The determiner of the base noun phrase is a first person possessive pronoun (`FPPP`), since we are interested in only the mentions of interviewee's immediate family members.
- The head noun of the base noun phrase is a family relation term (Table 5.2).



In the above example, the first person possessive pronoun *mon* is assigned `DET` by the part-of-speech tagger. It indicates that the following family relation belongs to the interviewee. On the other hand, the term *frère*, which is identified as a common noun (`NC`), is the *head* of this noun phrase. The adjective *grand* preceding the head noun further describes the head noun. The proper noun *Richard* that follows the head can be interpreted as name of elder brother of interviewee. Therefore, the phrase base noun phrase *Mon grand-frère Richard* is marked as the mention of an immediate family member of the interviewee. A few excerpts of varied base noun phrases (`BaseNP_DetNoun`) with mentions of the immediate or close family members of the interviewee are given in Table 5.3. These phrases are marked as `myFamily`.

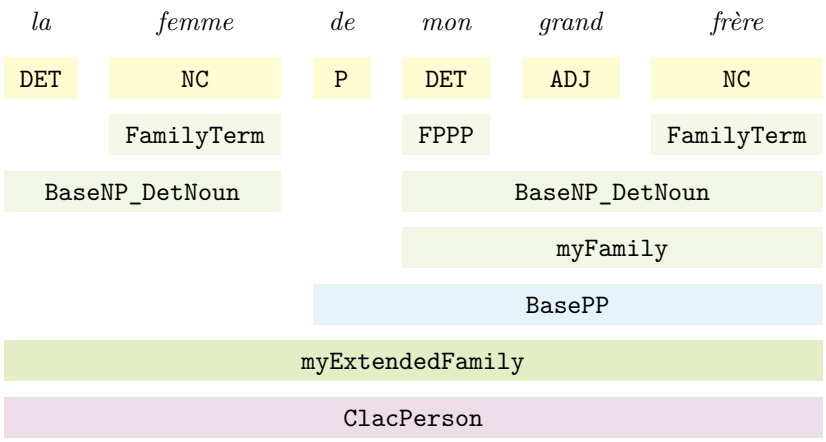
2. **Extended Family** While narrating their life-stories, the interviewees not only mention their immediate families, but also their extended family members. The mention of an extended family relations requires these three components in the given sequence -

   (a) **A base noun phrase mentioning a family relation** - A base noun phrase that contains a family relation from Table 5.2

   (b) ***du côté* (Optional)** - *du côté* in French means *"of side"* in English. Therefore, *du côté ma mère* in French means *on my mother's side* in English. This phrase is optional in the sequence of the components.

   (c) **A base prepositional phrase mentioning immediate family relation** - The final component in the sequence is a prepositional phrase where the head of the phrase is the preposition *de* and its complement is the base noun phrase referring to an immediate family relation (`myFamily`). The phrases extracted based on this rule are marked as `myExtendedFamily`.

   Below is an example of extended family relation extracted from the transcripts based on the above sequence -



The above example contains the phrase *mon grand frère* (my elder brother), that is a reference to interviewee's immediate family relation (`myFamily`). However, the context of this immediate family relation shows that it is not referring to the brother, but his brother's wife instead. Therefore to correctly capture the person mentioned in this excerpt, the grammatical context `myFamily` noun phrase is extended to the maximal length noun phase which starts at the first mention of the base noun phrase `BaseNP_DetNoun` containing mention of a family term from Table 5.2. Therefore, the head of this base noun phrase *femme*, is also the head of maximal length noun phrase *la femme de mon grand frère*. Since head defines the main concept talked about, this phrase talks about the wife of interviewee's elder brother and not his brother. The examples describes an instance of mention of an extended family member of interviewee. Table 5.4 gives a few such extracts.

les deux petites sœurs de ma maman
les deux tantes de Papa
Ma grand-maman du côté de ma père
mes tantes du côté de ma mère
deux grand-mamans du côté de mon père
le fils de mon oncle paternel

Table 5.4: A list of some phrases from the transcripts with the mentions of the member of the extended family of the interviewee.

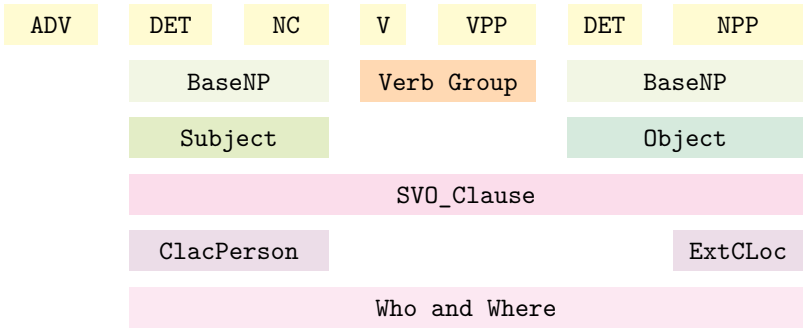
5.2.2 Connecting people and locations

In the beginning of the this chapter, we extracted SVO clauses where the utterance refers to the past. We have already identified the mentions of locations in the previous chapters and in the previous section, we have identified the mentions of people. Now we need to select a subset of the SVO clauses that connect people and locations that we have identified. It can be done by extracting the SVO clauses which contain both -

- `ClacPerson` - Mention of interviewee by first person pronoun, mention of interviewee's family relations or mention of the people by their names.
- `ExtCLoc` - Mention of a location

For example, let's consider the two sentences below -

1. *Alors* *mon* *frère* *a* *quitté* *le* *Rwanda*



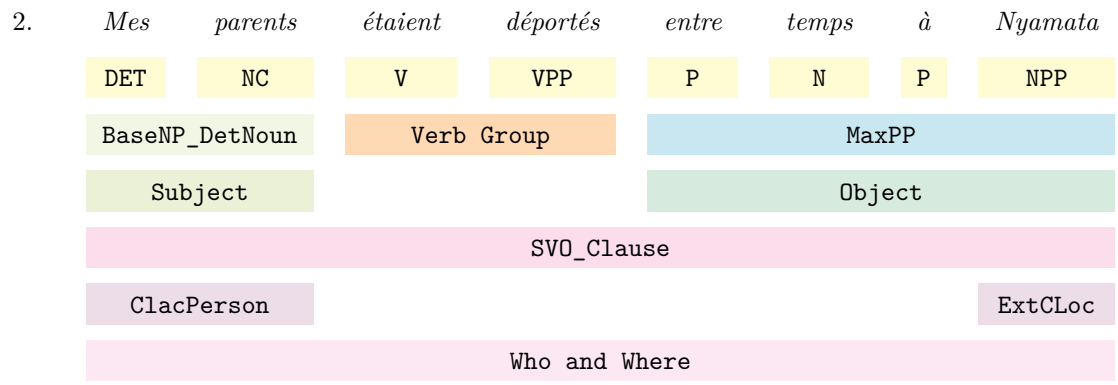


Table 5.5 gives a few samples from the transcripts where the mention of people was identified in connection to location mentions.

Mes sœurs ClacPerson *aussi étaient restées au Rwanda* ExtCLoc .
je ClacPerson *suis retourné au Rwanda* ExtCLoc *en soixante-neuf*
Je ClacPerson *suis allé chez les Frères Maristes à Bobandana* ExtCLoc
j' ClacPerson *ai travaillé au Nations Unies à Tripoli* ExtCLoc
je ClacPerson *suis revenu du Congo* ExtCLoc *en soixante-neuf*
Solange ClacPerson *sont nées au Congo* ExtCLoc

Table 5.5: A list of some `SVO_Clauses` from the transcripts where the people names are associated with locations

As we have seen in this chapter, based on a very simple grammatical analysis we can extract additional information related to the mention of locations. In this chapter, I have only discussed extraction of mention of people in association with location, especially in reference to past. However, it can also be used to identify other types of information, such as events, sentiment etc. By the use of this simple information extraction problem, this chapter only serves to emphasise the ability of such rudimentary grammatical analysis for extracting complex and inter-related pieces of information.

Chapter 6

Conclusion

In this thesis, the information extraction needs of researchers at the Geomedia Lab Concordia University are addressed through the CLaC Life Stories Pipeline with satisfactory results. In 2017, as state-of-the-art methods were not readily unavailable in our natural language processing environment, a simple grammatical approach detected location mentions in spoken text with reasonable accuracy comparable to human level annotations. The results of this research show that a simple grammar based approach is very close to the way humans would predict the mentions of a location in a text. Although, in this thesis the usefulness of grammar based approach to information extraction has been demonstrated for location detection, the objective of the thesis is not limited to this task. As discussed in the beginning of the introduction of this thesis, there are many real-world tasks that have complex information extraction requirements. The purpose of this thesis is to emphasize on the simplicity of using a grammar-based approach for such tasks. Moreover, in this thesis, we have discussed a particular use case of extracting complex related information, like the mentions of people related to interviewee, in reference to past and with relation to a location. Therefore, this thesis asserts that a simple grammatical analysis can be very useful in a lot of real world scenarios that have specific and complex information extraction needs where the state-of-art statistical approaches may not be directly applicable or even feasible.

References

Abney, S. (1997). "Part-of-Speech Tagging and Partial Parsing". In: *Corpus-Based Methods in Language and Speech Processing*. Springer Netherlands, pp. 118–136.

Albared, Mohammed, Marc Gallofré Ocaña, Abdullah Ghareb, and Tareq Al-Moslmi (2019). "Recent Progress of Named Entity Recognition over the Most Popular Datasets". In: *First International Conference of Intelligent Computing and Engineering*. IEEE, pp. 1–9.

Allen, James (1995). *Natural Language Understanding*. 2nd ed. Pearson.

Blache, Philippe and Azulay David-Olivier (2002). "Parsing Ill-formed Inputs with Constraint Graphs". In: *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 220–229.

Caquard, Sébastien and William Cartwright (2014). "Narrative Cartography: From Mapping Stories to the Narrative of Maps and Mapping". In: *The Cartographic Journal*.

Carlier, Anne, Michèle Goyens, and Béatrice Lamiroy (2013). "De: A Genitive Marker in French?: Its Grammaticalization Path from Latin to French". In: *The Genitive*. John Benjamins, pp. 141–216.

Chang, Yu-shan and Yun-Hsuan Sung (2005). "Applying Named Entity Recognition to Informal Text". In: *Recall* 1.

Chiticariu, Laura, Yunyao Li, and Frederick R Reiss (2013). "Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems!" In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 827–832.

Coates-Stephens, Sam (1992). "The Analysis and Acquisition of Proper Names for Robust Text Understanding". PhD thesis. City University London.

Crabbé, Benoit and Marie Candito (2008). "Expériences d'analyse syntaxique statistique du français". In: *15ème Conférence sur le Traitement Automatique des Langues Naturelles-TALN'08*, pp. 45–54.

Cunningham, Hamish, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, et al. (2014). *Developing Language Processing Components with GATE version 8 (A User Guide)*.

Eftimov, Tome, Barbara Koroušić Seljak, and Peter Korošec (2017). "A Rule-based Named-entity Recognition Method for Knowledge Extraction of Evidence-based Dietary Recommendations". In: *PLOS ONE* 12.6.

Ehrmann, Maud, Damien Nouvel, and Sophie Rosset (2016). "Named Entity Resources - Overview and Outlook". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pp. 3349–3356.

Erdmann, Alexander, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice Joyeux-Prunel, and Marie-Catherine de Marneffe (2019). "Practical, Efficient, and Customizable Active Learning for Named Entity Recognition in the Digital Humanities". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, pp. 2223–2234.

Goyal, Archana, Vishal Gupta, and Manish Kumar (2018). "Recent Named Entity Recognition and Classification Techniques: A Systematic Review". In: *Computer Science Review* 29, pp. 21–43.

Hadži, Vesna Požgaj, Damir Horga, and Tatjana Balazic Bulc (2012). "Speech fluency: a result of oral language proficiency?" In: *Linguistica* 52.1, pp. 87–100.

Kluegl, Peter, Martin Atzmueller, Tobias Hermann, and Frank Puppe (2009). "A Framework for Semi-Automatic Development of Rule-based Information Extraction Applications." In: *Proceedings of the Workshops on Learning, Knowledge Discovery, and Adaptivity (LWA)*, KDML–56.

Kluegl, Peter, Martin Toepfer, Philip-Daniel Beck, Georg Fette, and Frank Puppe (2016). "UIMA Ruta: Rapid development of rule-based information extraction applications". In: *Natural Language Engineering* 22.1, pp. 1–40.

Law, Jennifer H, Christopher Pettengell, Lisa W Le, Steven Aviv, Patricia DeMarco, David C Merritt, Sally CM Lau, Adrian G Sacher, and Natasha B Leighl (2019). "Generating Real-World Evidence: Using Automated Data Extraction to Replace Manual Chart Review." In: *Journal of Clinical Oncology* 37.

Li, Jing, Aixin Sun, Jianglei Han, and Chenliang Li (2020). "A Survey on Deep Learning for Named Entity Recognition". In: *IEEE Transactions on Knowledge and Data Engineering*.

Manning, Christopher D (2011). "Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?" In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pp. 171–189.

Marciano, Richard, William Underwood, Mohammad Hanaee, Connor Mullane, Aakanksha Singh, and Zayden Tethong (2018). "Automating the detection of personally identifiable information (PII) in Japanese-American WWII incarceration camp records". In: *Proceedings of the International Conference on Big Data*. The Institute of Electrical and Electronics Engineers, pp. 2725–2732.

Maynard, Diana, Wim Peters, and Yaoyong Li (2006). "Metrics for Evaluation of Ontology-Based Information Extraction". In: *Proceedings of the WWW Workshop on Evaluation of Ontologies for the Web*.

Milanova, Ivona, Jurij Silc, Miha Serucnik, Tome Eftimov, and Hristijan Gjoreski (2019). "LOCALE: A Rule-based Location Named-entity Recognition Method for Latin Text." In: *HistoInformatics@ TPDL*, pp. 13–20.

Nenadic, Goran, Irena Spasic, and Sophia Ananiadou (2003). "Terminology-Driven Mining of Biomedical Literature". In: *Bioinformatics* 19.8, pp. 938–943.

Palmer, David D (2000). "Tokenisation and Sentence Segmentation". In: *Handbook of Natural Language Processing*, pp. 11–35.

Poibeau, Thierry and Leila Kosseim (2001). "Proper Name Extraction from Non-journalistic Texts". In: *Computational Linguistics in the Netherlands 2000*. Brill Rodopi, pp. 144–157.

Richter, Ludwig, Johanna Geiß, Andreas Spitz, and Michael Gertz (2017). "HeidelPlace: An Extensible Framework for Geoparsing". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 85–90.

Toutanova, Kristina, Dan Klein, Christopher D. Manning, and Yoram Singer (2003). "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network". In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 252–259.

Vandeloise, Claude (1991). *Spatial Prepositions: A Case Study from French*. University of Chicago Press.

Wang, Ilaine, Aurore Pelletier, Jean-Yves Antoine, and Anaïs Halftermeyer (2020). "ODIL_Syntax: a Free Spontaneous Spoken French Treebank Annotated with Constituent Trees". In: *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 5301–5307.

Yadav, Vikas and Steven Bethard (2018). "A Survey on Recent Advances in Named Entity Recognition from Deep Learning models". In: *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2145–2158.

Appendix A

General Architecture for Text Engineering (GATE)

The General Architecture for Text Engineering (GATE) Framework (Cunningham et al., 2014) is an open-source platform that provides an integrated environment for linguistics processing and text analysis.

A.1 Basic Components

The GATE framework comprises of three basic components -

- **Language Resources (LRs) -** represent entities that form the data to be processed such as lexicons, corpora, and ontologies. The GATE supports various types of Language Resources to be analyzed in an automatic fashion. A set of documents called corpus, go through multiple processing phases such that the output annotation of each phase is available to the next phases in the pipeline.
- **Processing Resources (PRs) -** represent modules that process the text such as tokenizers, sentence splitters, parsers, etc. The output of each module is visualized in form of annotations on text spans.
- **Visualization Resources (VRs) -** consists of multiple visualization options and editing components that are useful for developing rules to capture text spans. For example, Figure A.1 shows a visualization of annotations in a document opened in GATE Developer. In this figure, the text spans highlighted in pink color are annotated as organized according to the ontology of annotations given in the right-hand side panel.

An application developed using the GATE Framework is a collection of the processing resources organized into sequential modules called the processing pipeline.

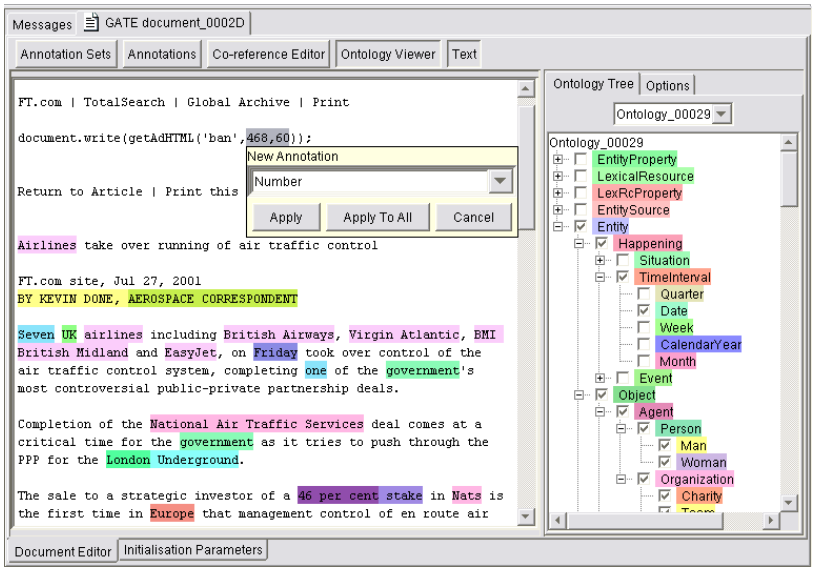


Figure A.1: A document opened in GATE Developer showing text spans highlighted for annotations

A.2 Processing Resources (PRs)

GATE provides a set of reusable processing resources for common natural language processing tasks. These resources are not definitive, and the user can replace and/or extend them as necessary.

A.2.1 Gazetteers

As described in GATE's online documentation[1] - A gazetteer in GATE consists of a set of lists containing names of entities such as drug names, symptoms, units of measurements, person name titles, location names, etc. These lists are used to find occurrences of these names in text. When a gazetteer processing resource is run on a document, annotations of type Lookup are created for each matching string in the text. The gazetteer list does not require sentences or tokens to be identified on text as it finds matches based on the names in the list and corresponding text spans in the respective document. This makes them a very useful tool to identify entities in text. In our approach, we do not use an existing list of location names to identify location name mentions in transcripts. Instead, we use a grammar-based methodology to identify mentions of location names in text. Next, I define a freely available module in GATE called the gazetteer list collector[2] that is complements our grammar-based approach to detect location names automatically.

A.2.2 The Gazetteer List Collector

The Gazetteer List Collector is a module in GATE that provides the facility to collect the text spans against an annotation into a list. For example, the `CLSP` utilized this tool to collect all occurrences

[1] https://gate.ac.uk/sale/tao/splitch13.html#sec:gazetteers:intro
[2] https://gate.ac.uk/sale/tao/splitch13.html#sec:gazetteers:listscollector

of location names identified in text as `ClacLoc`. The gazetteer list collector populates a gazetteer list with these names.

A.2.3 Tokenizers

GATE has a variety of tokenizers, like GATE English Tokenizer, English Tweet Tokenizer, and Unicode Tokenizer, that splits text into simple tokens, such as numbers, punctuation marks, symbols, and words of different types using a set of regular expressions. Each Token annotation has a feature called `kind` whose value can be a number, punctuation mark, symbol, or a word. In the case of `word`, the information concerning the orthography of the word is also retrieved and saved as a feature `orth`.

A.2.4 Sentence Splitter

The sentence splitter segments the text into sentences, e.g. GATE Sentence Splitter, RegEx Sentence splitter, etc. The GATE RegEx Sentence Splitter uses regular expressions based on the syntactic rules for sentence identification.

A.2.5 Part-of-Speech Tagger

Another important module in the natural language processing pipeline is the part-of-speech tagger. The part-of-speech tagging module identifies the grammatical class of each word and assigns it a part-of-speech tag. Examples include Tree Tagger, Stanford POS Tagger (with different language models), and GATE Twitter POS Tagger.

A.2.6 Java Annotation Pattern Engine (JAPE) Transducer

The Java Annotation Pattern Engine (JAPE) is a component of GATE. It is based on regular expressions and therefore the text is read by patterns from left to right, and pattern boundaries are extended incrementally in a monotonic manner. A `JAPE` grammar consists of a set of phases that are independent code segments. Each of the phases consists of a set of pattern/action rules that run sequentially. Patterns can be specified by describing a specific text string, or annotations previously created by modules such as the tokenizer, gazetteer, or parser. The patterns captured with JAPE rules have an output in linear time as it annotates patterns on the text using finite-state pattern-action rules. This allows a very fast and efficient way of creating new annotations on text if specified patterns are found. Hence the `JAPE` grammar consists of a set of rules to annotate the text for certain patterns.

A.3 ANNIE: A Nearly-New Information Extraction System

ANNIE (A Nearly-New Information Extraction System) is an information extraction available as a default component of GATE. A processing pipeline is available for processing text using ANNIE and

it consists of the following main processing resources: a tokenizer, sentence splitter, POS tagger, gazetteer, and finite-state transducers (based on GATE's built-in language JAPE that allows regular expressions over text annotations). Although the default IE pipeline ANNIE is for processing text in the English language, it has been adapted for the task of IE in other languages. ANNIE plugins are available in multiple natural languages including French, German, Urdu, Persian, Arabic, Chinese and Russian. The GATE's ANNIE adaptation to the French Language is called `FRENCH NE`. It is a baseline Named Entity Recognition pipeline for NER on French texts.

- **The Unicode Tokenizer** for tokenization
- **The RegEx Sentence Splitter** for sentence splitting
- **The Stanford's parts-of-speech tagger** v3.8.0[3] is used for NER on French text by the FRENCH NE processing pipeline. The Stanford POS Tagger adds a `category` feature to the `Token` annotation created by the GATE's Unicode Tokenizer. Table A.1, gives tagset of Stanford POS Tagger (Wang et al., 2020). The Stanford POS Tagger is trained on French Tree Bank's CC tag-set (Crabbé et al., 2008). The pos-tagger has a slightly different tag set than the language treebank.

Table A.1: Stanford POS Tagger tagset

Stanford tags	lexical category	sample tokens from transcripts
ADJ	adjective	important
ADJWH	interrogative adjective	Quel, quelle, quels, quelles
ADV	adverb	pas, tout, aussi, bien, non
ADVWH	interrogative adverb	comment
CC	coordination conjunction	et, mais, ou
CLO	object clitic pronoun	le
CLR	reflexive clitic pronoun	m'
CLS	subject clitic pronoun	J', ils, Elle, On
CS	subordination conjunction	comme, quand
DET	determiner	mes, cette, mon, un, trois
DETWH	interrogative determiner	quel before noun
ET	foreign word	euh, Umurwanashyaka
I	interjection	wow, Boum
NC	common noun	partie, heure, groupe, routes
NPP	proper noun	Congo, Angleterre, Canada
N	abstract or mass noun	moment, sabotage, Logiciel
P	preposition	en, au, dans, sur, avec, des
PUNC	punctuation mark	; ? :
PREF	prefix	Saint-
PRO	full pronoun	eux
PROREL	relative pronoun	qui
PROWH	interrogative pronoun	qui , que
V	indicative or conditional verb form	ai, a, voudraise
VIMP	imperative verb form	arrêtez, montez
VINF	infinitive verb form	poser, dire, voyager
VPP	past participle	fait, perdu
VPR	present participle	voulant, donnant
VS	subjunctive verb form	puissent

[3]https://nlp.stanford.edu/software/stanford-postagger-full-2017-06-09.zip

- **NE Jape Transducers** The Named Entity Transducers to detect mentions of person names, location names, money, percentages, and dates using JAPE rules and gazetteers lists of names.