FRACAS: a FRench Annotated Corpus of Attribution relations in newS

Richard, Ange (1, 2) Alonzo Canul, Laura C. (1) Portet, François (1)
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
(2) Univ. Grenoble Alpes, CNRS, Sciences Po Grenoble, Pacte, 38000 Grenoble, France
{ange.richard, laura.alonzo-canul, francois.portet}@univ-grenoble-alpes.fr

Abstract

Quotation extraction is a widely useful task both from a sociological and from a Natural Language Processing perspective. However, very little data is available to study this task in languages other than English. In this paper, we present FRACAS, a manually annotated corpus of 1,676 newswire texts in French for quotation extraction and source attribution. We first describe the composition of our corpus and the choices that were made in selecting the data. We then detail the annotation guidelines, the annotation process and give relevant statistics about our corpus.We give results for the inter-annotator agreement which is substantially high for such a difficult linguistic phenomenon. We use this new resource to test the ability of a neural state-of-the-art relation extraction system to extract quotes and their source and we compare this model to the latest available system for quotation extraction for the French language, which is rule-based. Experiments using our dataset on the state-of-the-art system show very promising results considering the difficulty of the task at hand.

Keywords: attribution,relation extraction,corpus

1. Introduction

Automatic quotation extraction and source attribution, namely finding the source speaker of a quote in a text, is a widely useful, however overlooked, task: it has many applications, both on a Social Science perspective (for example, for fact-checking, detection of fake news or tracking the propagation of quotes throughout news media) and on a Natural Language Processing perspective (it can be tackled as a text classification task or a relation extraction task, as well as entail coreference resolution). It is however a complex task, both to define and to solve, and as such has not been widely researched in NLP. There is little available corpus in English (Pareti, 2012; Papay and Padó, 2020; Vaucher et al., 2021), and none for French, which is the language we aim to study here.

In this article, we contribute to the study of quotation extraction and source attribution in two ways:

1. We make available FRACAS, a human-annotated corpus of 10,965 attribution relations (quotes attributed to a speaker), annotated over a set of 1676 newswire texts in French. This corpus contains labelled information on quotations in each text, their cue and their source, as well as the speaker's gender. Details on how to request the data can be found in Section 8.

2. We describe a set of experiments using our corpus with a relation extraction system, which we compare to our baseline for quotation extraction for French, a rule-based system developed in Soumah et al. (2023).

3. We show that training more recent architectures of relation extraction for this task with our corpus substantially improves the current results for French quotation extraction.

We begin this paper by discussing different definitions of our task, related work and existing corpora in Section 2. In Section 3, we give detailed information and statistics about FRACAS. We follow by detailing the annotation process in Section 4: we describe our annotation guidelines, annotation campaign and results for inter-annotator agreement. We finish with Sections 5 and 6 by describing experiments and results using our corpus on two systems, a rule-based system and a state-of-the art model on relation extraction.

2. Related Work

2.1. Task Definition

Quotation is not a straightforward linguistic phenomenon, and has not been widely studied for French, although there has been a renewal of interest for its study in recent years. There is thus very little available corpora to work with, and no consensus on what the task of quotation extraction entails. The task understood as *quotation extraction* aims to detect text spans that correspond to the content of a quote in a text. This task, although a seemingly straight-forward one, is deceptively simple: a quote might be announced by a cue, and may or

LREC-COLING 2024, pages 7417–7428

may not be enclosed between quotation marks – sometimes, it also contains misleading quotation marks. Its span can be very long and discontinuous, or overlap with a cue element. Quotations are usually divided into three types: direct (enclosed in quotation marks), indirect (paraphrase), and mixed or partially indirect (a combination of both). We describe these types in detail in Section 4. All these types of quotations can be found indiscriminately within different types of texts, in literary texts or in news documents, which makes their automatic identification all the more difficult.

A subsequent challenge to quotation extraction is the identification of the quoted speaker: this task is known as *source attribution*. It involves identifying for each quotation its source entity. The goal is thus not only one of sequence classification (**quotation extraction**) but also one of relation extraction (**source attribution**) if it also involves identifying the speaker of each quote.

2.2. State of the art

The literature contains many different takes on this same task, as described in Scheible et al. (2016) and Vaucher et al. (2021)'s states of the art. There are different levels of granularity to tackling this phenomenon. Some works define quotation extraction as a sentence classification task: in Brunner (2013), for example, the aim is to train a classifier to identify if a sentence input contains a quote, without emphasis on detecting the boundaries of the quote in itself.

Other works look at the task as one of text sequence classification: they focus on extracting quote content as entities from within a text. Some of these works only consider direct quotes, which are much easier to detect due to the almost constant presence of quotations marks surrounding them. Other works adopt a wider definition of quotations and include indirect and mixed quotes. Contributions in this line largely adopt a rule-based approach using patterns to identify quotes, speech verbs gazetteers and syntactic pattern recognition (Pouliquen et al., 2007; Salway et al., 2017; Soumah et al., 2023). It has to be noted that some works choose to apply neural architectures to detect quotations: Scheible et al. (2016) use a pipeline system that first detects cues then quotations and combines a perceptron model and a semi-Markov model, while Papay and Padó (2019) use an LSTM-based approach to detect quotation spans.

More complex approaches consider the task as a relation extraction task and seek to not only detect quote entities and cues, but to link these to their speaker entities. Up until recently, the state-of-the-art system for quotation extraction for English was the one developed by Pareti (2015), which uses a pipeline system: it first extracts cue entities with a k-NN classifier, then uses a linear-chain conditional random field (CRF) to extract quotation spans in the close context of each cue. The results are then used as an input to a logistic regression speaker attribution model developed in O'Keefe et al. (2013).

Our own approach is most similar to works inspired by Pareti (2015), in the sense that we consider all types of quotations (direct, indirect and mixed) and seek to perform both the task of quote extraction and of source attribution. Papay and Padó (2019) describe several systems with a similar goal. These systems are for the most part either rule-based or neural network-based, are trained on English data, and predate the development of BERT-like large language models which are now widely used to solve a various range of NLP tasks.

As for many complex NLP tasks, only few systems exist for other languages (Tu et al. (2021) for German, Salway et al. (2017) for Norwegian, Sarmento and Nunes (2009) for Portuguese). For French, existing work is very scarce. Previous work on automatic quotation extraction for French date back from more than a decade ago and are mostly systems based on syntactic rules and lexicons (Pouliquen et al., 2007; Poulard et al., 2008; De la Clergerie et al., 2009; Sagot et al., 2010).

The scarcity of systems tailored to quotation extraction tasks leads us to consider our target task from a wider perspective. As stated previously, quotation extraction can be defined as a relation extraction task, which has been more widely covered in the literature. Relation extraction is a difficult task to solve: Yao et al. (2019) were holding the state-of-art on document-level relation extraction with a F1 score of 51.06% before the spread of the use of Transformers architectures. As it is, later works based on the use of pretrained large language models were able to obtain better performances in recent years, as shown in Table 1 for benchmark datasets. These systems do not necessarily follow the same architecture, but all make use of pre-trained large language models like BERT or GPT. Many of these systems do not adopt previous approaches that follow a two-step process consisting of first sequence classification then relation extraction. Tackling both aspects in a parallel fashion aims at mitigating error propagation. It has to be noted however that even with recent advances in performances, state-of-the-art scores for relation extraction remain only moderately high, especially for document-level relation extraction, which is our target task here.

2.3. Available corpora

Since quotation extraction is not one of the most explored NLP tasks, very few manually labelled

Dataset	F1	Reference
TACRED	76.80	Wang et al. (2022)
CoNLL04	76.65	Cabot and Navigli (2021)
ACE2005	73.00	Ye et al. (2022)
DocRED	67.53	Ma et al. (2023)
RED^{FM}(Fr)	52.50	Cabot et al. (2023)

Table 1: State-of-art results on different relation extraction datasets (all sentence-level relations except for DocRED which is document-level relations)

corpora are available for training and evaluation.

The PARC3 English corpus (Pareti, 2012) is one of the early corpora for quote extraction. It comes as an additional layer to the Penn TreeBank corpus, which is not freely available itself. Since PARC3, more recent corpora have been released for English. For instance, PolNeAR v.1.0.0 (Newell et al., 2018) is composed of articles by 7 U.S.A. national news outlets covering the U.S.A. General Election campaigns of 2016. These 1,008 articles are annotated with source, cue and content labels. We can also mention the SUMREN benchmark (Gangi Reddy et al., 2023) which contains 745 texts from 4 news sources that contain reported speech annotations, but no annotation of relations. To our knowledge, the largest existing dataset is Quotebank (Vaucher et al., 2021), a corpus of 178 million articles from the Spinn3r news corpus containing automatic annotations of quotes. The annotation was done using Quobert, a BERT-based model designed by the authors to extract direct and indirect quotations as well as perform speaker attribution. To our knowledge, the RiQuA corpus (Papay and Padó, 2020) is the only freely available corpus which has been manually annotated for quotation extraction and source attribution. However, it only focuses on literary texts in English.

In languages other than English, corpora become scarce. We can mention the RWG corpus (Brunner, 2013), a collection of German narrative text from the 1787–1913 period annotated in direct, indirect, free indirect, and reported variants of speech. Zulaika et al. (2022)'s sentence classification corpora for Spanish and Basque are also available, but do not contain information about speakers, cues and quote boundaries. As for French, we were not able to find any freely available labelled corpus on French quotations.

In this article, we describe FRACAS, the first freely available corpus for quotation extraction and source attribution for French. The corpus contains 10,965 attribution relations over 1,676 newswire texts. Labelled entities include direct, indirect and mixed quotations. Each is linked to entities corresponding to their source speaker and, optionally, to a cue that introduces the content of the quote. We chose newswire texts as they are more likely to contain many examples of quotations, as journalistic writing is most often based on pieces of reported speech that the journalist collected during their reporting (Nylund, 2003). This corpus also contains coreference annotations for quotation speakers – namely, when the source speaker is a pronoun. We detail the contents of the corpus and the annotation process in the following section.

3. Our corpus

To be able to produce an annotated corpus of French news articles, our first task was to find a corpus free to use and to redistribute. There are several French news articles or newswire corpora available for research, but most of them are very lightly documented as to their origin. Other better documented corpora are not free nor available to redistribute. We chose to use the Reuters Corpora Reuters-21578, Distribution 1.0, a multilingual corpus of newswires from the British news agency Reuters. These newswires were published between 1996, August 20th and 1997, August 19th, and the corpus was made available in 2005. The multilingual version contains 487,000 newswires written in 13 different language by local journalists – it was not produced by automatic translation. This corpus is freely distributed upon request by the National Institute of Standards and Technology (NIST, 2005) of the United States and is originally used for document classification tasks.

Our goal was to produce around 1,500 annotated documents, each document annotated by two annotators. 1,500 documents were thus randomly picked from the total 85,710 documents of the French part of the corpus. We applied on each drawn document our baseline system for quotation extraction, a rule-based algorithm by Simon Fraser University's *Discourse Lab* (Soumah et al., 2023) that we present in more details below, to make sure that the final corpus was only made of documents containing at least one quote.

Another batch of documents was later added to our original corpus, as after the first round of annotation, we observed that the gender ratio between quotes by men and quotes by women was highly unbalanced. This was a problem as our end goal for this task is to use our quote extraction model to measure gender imbalance in the news. We chose to pick another 160 files to raise the number of quotes by women. The final gender ratio is unfortunately still far from being balanced, as shown

Partition	#docs	#tokens	Mean #tok. per doc
train	1,114	436,150	391.51
dev	281	131,202	466.91
test	281	137,083	487.83
TOTAL	1,676	704,435	448.75

Table 2: Number of documents and tokens in the FRACAS

in Table 4.[1] The low presence of quoted women in the newswires might be explained by the fact that in the mid-1990s, the presence of women was even rarer in media than it is today.

The final corpus contains 1,676 documents, divided into train, development and test sets as shown in Table 2. The splitting process into sets is detailed in section 5.1.

4. Annotation Process

This section is dedicated to the annotation process that lead to FRACAS: we first detail the annotation guidelines, which accounts for all entity and relation labels that can be found in the final corpus. We describe the manual annotation campaign then give results for inter-annotator agreement, which are fairly satisfying given the complexity of our task.

4.1. Annotation Guidelines

We draw on Pareti (2012)'s work for the annotation guidelines and labels. We consider a quote as a triplet made of three entities: a quote content, linked to a speaker by a *Quoted in* relation, and linked to a cue by an *Indicates* relation (this last entity is optional, but most quotes are introduced by a cue, very often a verb). We distinguish quote types and speaker types. Quote types are the following:

- **Direct Quotation**: a direct quotation reports the quoted speaker's exact words. It is the easiest type to spot as it is usually enclosed by quotation marks.

  (1) [Nicki]$_{\text{SPEAKER}}$ [said]$_{\text{CUE}}$ [“Let's go to the beach!”]$_{\text{QUOTE}}$.

- **Indirect Quotation**: an indirect quotation is a paraphrase of the speaker's words. It is usually a rephrasing of these words, and is most often written in the 3rd person, without quotation marks.

  (2) [Rihanna]$_{\text{SPEAKER}}$ [asked]$_{\text{CUE}}$ [not to stop the music]$_{\text{QUOTE}}$.

- **Mixed Quotation**: a mixed quotation is a paraphrase (indirect quotation) that contains direct speech elements (words or part of a sentence), usually enclosed within quotation marks.

  (3) [Britney]$_{\text{SPEAKER}}$ [said]$_{\text{CUE}}$ that [she did it “again”]$_{\text{QUOTE}}$.

We divide Speaker labels as following: **Agent** (when the speaker is a single person, i.e “Mariah Carey”), **Group of People** (i.e “The cast of *Drag Race France*”), **Organization** (i.e “the UNESCO”), or **Source Pronoun** (i.e “she”). In that last case, the pronoun is linked to its referent entity, labelled with one of the Speaker tags, as in the following example.[2] Co-reference relations are only indicated for Speaker pronouns which are related to Quote entities, and not on all pronouns found in the text.

(4) [Beyoncé]$_{\text{SPEAKER (Agent)}}$ warned him! [She]$_{\text{SPEAKER (Source pronoun)}}$ [told]$_{\text{CUE}}$ him [he should have put a ring on it]$_{\text{QUOTE}}$.

Additionally, each speaker is labelled with a gender tag amongst the following: Male, Female, Mixed, Other or Unknown. The gender was assigned based on linguistic features like gender agreement and other semantic references that could be found within the text. Each quote is linked by a relation to a speaker and a cue, and each source pronoun is linked to a referent labelled with one of the above speaker labels. The overall numbers of tagged entities for each label is detailed in Table 3.

4.2. Annotators

We used the software BRAT (Stenetorp et al., 2012) for this annotation task, as shown in Figure 1. The original corpus (1,436 newswires) was annotated by a team of 9 annotators: 7 women, 1 men and 1 non-binary person, all graduate students from NLP, Communication Studies or Linguistic degrees recruited through an open call through the university channels. The annotators

[1] An additional ”Other” gender tag was also available for entities whose gender did not fall in the Male/Female binary. This was originally intended for cases of non-binary Agent speakers, but these were absent from our corpus, most likely due to the date of the documents. This label ended being used only 13 times by annotators in the whole corpus, all to tag ambiguous cases of non-human Agent entities like “un sondage” (“a poll”) or ”les premiers pas de l'enquête” (“the detective's first findings”). We chose to remove them from this table.

[2] Note that in this example, the first sentence is not considered as a quote, even with the presence of the ambiguous speech verb “warned”.

	QUOTES			SPEAKER (as direct source / as coreferent)								CUE
	Direct	Indirect	Mixed	Agent		Organization		GoP		SP		
train	2,698$^{(40)}$	2,919$^{(43)}$	1,123$^{(17)}$	1,881$^{(34)}$	1,196$^{(91)}$	1,071$^{(20)}$	83$^{(6)}$	628$^{(11)}$	38$^{(3)}$	1,928$^{(35)}$		6,439
dev	820$^{(40)}$	835$^{(41)}$	404$^{(19)}$	570$^{(40)}$	312$^{(84)}$	386$^{(27)}$	43$^{(11)}$	158$^{(11)}$	17$^{(5)}$	312$^{(22)}$		1,978
test	919$^{(42)}$	918$^{(41)}$	382$^{(17)}$	628$^{(41)}$	353$^{(86)}$	360$^{(24)}$	36$^{(9)}$	191$^{(12)}$	21$^{(5)}$	353$^{(23)}$		2,126
TOTAL	4,437$^{(40)}$	4,672$^{(43)}$	1,909$^{(17)}$	3,079$^{(41)}$	1,861$^{(89)}$	1,817$^{(25)}$	162$^{(8)}$	677$^{(9)}$	76$^{(3)}$	1,861$^{(25)}$		10,543

Table 3: Number *(and % in partition)* of annotations per entity in FRACAS (GoP = Group of People, SP = Source Pronoun)

	Ma.	Fe.	Mix.	Unk.
train	2,706$^{(79)}$	257$^{(7)}$	132$^{(4)}$	336$^{(10)}$
dev	613$^{(58)}$	246$^{(23)}$	92$^{(9)}$	101$^{(10)}$
test	705$^{(60)}$	254$^{(21)}$	110$^{(9)}$	117$^{(10)}$
TOTAL	4,024$^{(71)}$	757$^{(13)}$	334$^{(6)}$	554$^{(10)}$

Table 4: Gender distribution of speaker entities *(and % in partition)* for Agents and Group of People only.

were paid according to the French minimum wage (a gross salary of €18.75 per hour) for a 20-hour-long contract, and each had to annotate 300 documents. The documents were split by batches of 150 and each batch was annotated by 2 different annotators. The annotators received detailed annotation guidelines and half a day of online training with the annotation campaign supervisor to make sure that the guidelines and the use of the software were understood. The annotators had about a month to complete the annotation task, between June and July 2021. At halfway point, the ongoing annotation was checked by the supervisor and individual feedback was sent to annotators to clarify misunderstood instructions.

The additional 168 documents that were later added to raise the number of quotes by women were annotated by three expert annotators: two researchers from the project and one of the annotators previously trained for the campaign. For this subsequent annotation campaign, the documents were divided amongst the three annotators, with an intersection of 30% of the documents that were annotated by all three annotators to calculate inter-annotator agreement.

4.3. Inter-Annotator Agreement (IAA)

After both annotation campaigns, the annotated documents were cleaned and preprocessed before computing IAA to correct some easy-to-fix annotation errors. For instance, we automatically edited annotated spans that included wrong boundaries such as extra punctuation or white spaces, or were missing elements, such as direct quotes entities that did not include their quotation marks. We also noticed that some guidelines had not been understood by all annotators: the instructions were to annotate any elements that were syntactically linked to a speaker (i.e. in a speaker phrase like “Madonna, queen of pop”, the whole phrase should be annotated as a Speaker and not only “Madonna”). However, some annotators did not annotate all the elements. We chose to reprocess these instances by keeping the longest version of an annotated phrase if two Speaker entities were overlapping between two annotators.

To measure inter-annotator agreement for entity annotation, we use the γ score proposed by Mathet et al. (2015). Since our annotation task was one of sequence delimitation and classification, the γ score allows us to compute an agreement that accounts for both unitizing (agreement on unit span location within a text) and categorization (agreement on unit labeling). The γ score, like Cohen’s κ, is computed from observed and expected disagreements, instead of agreements. We consider it the best IAA score for our task as it takes into account overlap when calculating unit alignment, chance correction and category weight (disagreements on rare categories are more serious than on frequent ones). The results for all entities is showcased on Table 5. We obtain an overall inter-annotator agreement score of 0.77, which is satisfying considering the difficulty of the task at hand. We observe the best IAA scores for Direct Quotation and Source pronoun entities, as well as for Cues and Agent speakers. Indirect Quotation annotations obtain the lowest score. We link this to the difficulty to determine what is a paraphrase or not, and what is to be included in the span of the quotation. These score allow us to note that the quotation detection task, when extended to its larger comprehension (meaning including indirect and mixed quotations), is not an easy one even by human standards.

To determine which annotations should be included in the final corpus, we computed an IAA with relation to a gold standard for each batch of 150 document annotated by a pair of annota-

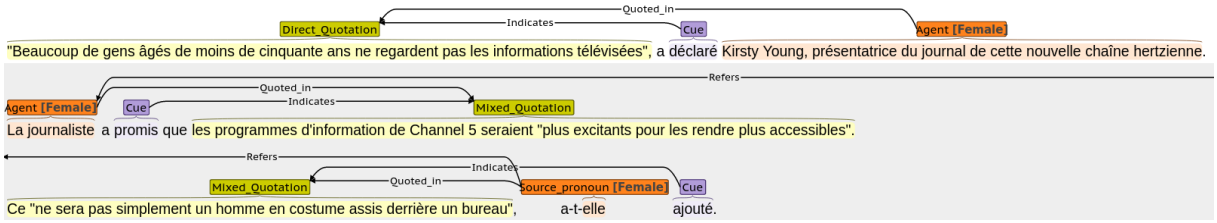

Figure 1: Screenshot of an annotated text in the BRAT interface

Entity label	γ agreement
All entities	0.7699
Direct Quotation (Q)	0.8857
Indirect Quotation (Q)	0.6415
Mixed Quotation (Q)	0.7468
Cue	0.8291
Agent (S)	0.8337
Organization (S)	0.7858
Group of people (S)	0.7828
Source pronoun (S)	0.8980

Table 5: γ agreement between annotators of first annotation campaign (S = Speaker entity types, Q = Quotation entity types)

tors. The gold standard was composed of 10 documents of each batch annotated by an expert annotator. An IAA score was computed with the same γ measure for each annotator for each batch over the 10 documents. The documents annotated by the annotator who had the highest IAA with the gold standard were then kept in the final corpus.

5. Experimental Setup

We present in this section experiments using our corpus. We start by detailing the necessary splitting and pre-processing of our data. We then describe our evaluation criteria for relation extraction before continuing with the description of our tested systems, a rule-based system and a generation model.

5.1. Data Pre-processing

Splitting the data. We divide our corpus into train (66%), development (13%) and test (13%) sets as shown in Tables 2, 3 and 4. Aware of the relatively small size of our dataset, the complexity of the task and the sample imbalance in speaker and quote labels, we divide documents among the data splits based on a set of different criteria (see listing below). More specifically, we tag each document with the criteria they meet from the list (following the exact same order of priority) and divide groups of documents meeting the same criteria using the desired train/dev/test ratios (in the case of a shortage, we satisfy numbers for dev and test; in the case of a surplus, additional samples are assigned to train). This allows us to fairly distribute samples with underrepresented labels among the splits, but also to create particularly difficult dev and test sets for evaluation.

1. The document includes at least one quote by a female Agent
2. The document includes at least one quote by a mixed-gendered group of speaker
3. The document includes at least one quote by a speaker who is a Group of People or an Organization (i.e: not an Agent)
4. The document includes at least one Mixed Quotation
5. The document includes at least one Indirect Quotation
6. The length of the document is in the 95th percentile

Paragraph building. Newswire texts have the characteristic of widely varying in length: while some documents might contain a single short paragraph summarizing an event, others might be composed of several (short or long) sub-paragraphs of detailed reporting. Training a neural model with long documents is often a difficult task. However, news data offer an advantage over other textual resources: because it is meant to be accessed and digested quickly, information across paragraphs tends to be well isolated, and information within paragraphs tends to be well compacted. We take advantage of these features and build sub-documents out of the "paragraphs" from each document by merging spans of texts delimited by the newline character until an annotated relation is found. For relations that are spread over multiple newline-delimited spans of text, (as is often the case of *Refers* relations), we keep the relation if all its components can be found across the current or previous two spans. We remove it otherwise. We find that with this technique, we are able to create

documents that do not cross the 512 token mark and remove only 1% of all annotated relations.

5.2. RE Evaluation

We evaluate our models on RE (Relation Extraction) using Precision, Recall and Micro-F1 scores as showcased in Table 6. We report results for both "strict" and "boundaries" evaluation modes. In a "strict" evaluation, a relation is considered correct if entity boundaries, entity labels and relation label match the gold standard, whereas in a "boundaries" evaluation, only correct entity boundaries and relation label are required.

5.3. Modeling Approaches

5.3.1. Baseline: *Radar de Parité* (Soumah et al., 2023)

The latest (and only) freely available system for quotation extraction and source attribution for French is *Radar de Parité* (Soumah et al., 2023), a syntactic rule-based quote extractor developed for the *Gender Gap Tracker* project, which aims at measuring gender inequalities in reported speech in Canadian news written in French. The *Radar de Parité* was designed to extract direct and indirect quotes from texts as well as their speaker, making it therefore a suitable baseline for our task. We refer the reader to Soumah et al. (2023) for an in-depth description.

Differences with FRACAS A notable difference between the *Radar de Parité* and our data is that the considered entities for a quotation triplet differ slightly: the *Radar de Parité* extracts quotes, speakers and cue but does not distinguish between quote and speaker subtypes as we do. To evaluate their system on our data, we chose to compare entity types according to overall category tags and not on subtypes tags: our Direct, Indirect and Mixed quotations were considered as "quotes" and our Agents, Organizations, Groups of People, Source Pronouns were considered as "speakers". We also chose not to evaluate co-reference relations, as no reference for pronouns were given in the output by the system.

Evaluation We evaluate *Radar de Parité* on our dev set with a focus on relation extraction. The main shortcoming of this rule-based system is that it does not detect exact entity boundaries. As a result, a strict evaluation of relation extraction that takes into account exact span match for linked entities yields extremely poor results. This is why we also evaluated this system while allowing a margin on entity boundaries: for each predicted relation composed of two entities and a relation label, entities are considered as correct if the predicted entity spans overlap with at least either 30% (loose agreement) or 80% (strong agreement) of the gold entity spans. This allowed margin is the one used by Soumah et al. (2023) in their own evaluation.

Using the strict mode of evaluation described in Section 5.2, we obtain a micro F1 score of 52.59 for a minimum of 80% overlap between predicted and gold standard spans, and a micro F1 score of 62.66 for a minimum of 30% overlap. The system performs better on *Indicates* relations between cues and quotes (with a 57.01 F-score for a 80% overlap and a 67.68 F-score for a 30% overlap) than on *Quoted in* relations which attribute a speaker to a quote (with a 48.48 F1 for 80% overlap and a 58.74 F1 for 30% overlap).

While aware that the differences in labelling and task focus do not make the *Radar de Parité* system directly applicable to our data, we hold it as a baseline to evaluate how a more advanced NLP model could improve performance, particularly with relation to extracting exact entity boundaries. Our task deals with both long (namely: quotes) and short (cues) entities, which constitutes one of its challenges.

5.3.2. Quotation Extraction as a Generation Task: *REBEL* (Cabot and Navigli, 2021)

We model quotation extraction and source attribution as a generation task using *REBEL* (Cabot and Navigli, 2021), a framework in which relation extraction is re-framed as a sequence to sequence task and then solved with an autoregressive generation model based on BART-large (Lewis et al., 2020). The input and output for *REBEL* are the text containing the relations in its raw form and the linearized relation triplets to be decoded by the model, respectively. In our experiments, we make use of $mREBEL_{32}$ a multilingual version of REBEL fine-tuned on 32 relation types and 18 languages, including French. We refer the reader to Cabot et al. (2023) for further details on $mREBEL_{32}$'s architecture and training data.

Triplet linearization. Dealing with quotes when building the input and output for $mREBEL_{32}$ poses two main challenges. First, quotes tend to be long texts that are sometimes split across sentences. Hence, a clever way of linearizing the data must be devised: for one, as to avoid generating an output that will be too long for pre-trained models to process (most pre-trained models for the French language are constrained to process inputs with a maximum of 1,024 tokens length), and for another, as to be able to express the relations within the text in a way that is not too hard for the model to decode. Second, such

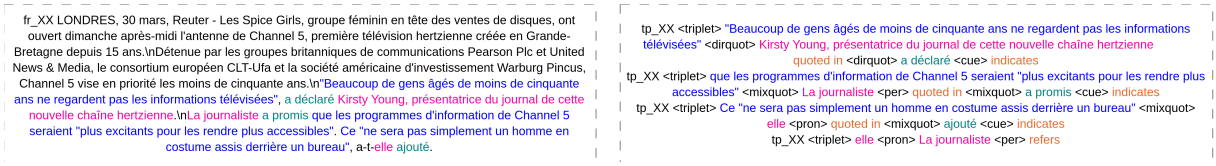

Figure 2: Example input and output built for a single document (newswire text) in our dataset. the boxes on the left and right show the raw text and the linearized triplets from the text, respectively.

System	QI	I	R	Prec	Rec	F1
$mREBEL_{bo}$	66.21	70.83	43.24	70.59	65.15	67.76
$mREBEL_{st}$	62.07	69.09	43.24	67.60	62.40	64.89

Table 6: Relation extraction results on the dev set of FRACAS. Relation scores per relation type (QI: quoted in, I: indicates, R: refers) are given.

linearization must take into account the fact that referent information (contained and modeled by *Refers* relations), as explained in Section 4, might not always be available. To tackle these challenges, we linearize our triplets similarly to Cabot and Navigli (2021) but with some key adjustments. First, we do not group triplets by head but by tail entity, which in our data always corresponds to a quote or a speaker (and more precisely, a source pronoun, see Section 4). Second, we build separate triplets for triples holding a *Refers* relation, which allows us not only to compact the information within quote triplets but also to easily dispose of referent information for training when they are scarce. Finally, we also add new special tokens for each of our entity tags. An example of the final training data for $mREBEL_{32}$ is shown in Figure 2.

Implementation details Models are implemented using PyTorch Lightning (Falcon and The PyTorch Lightning team, 2019) and Optuna (Akiba et al., 2019) for hyper-parameter optimization. Training is done over a single NVIDIA A100 80GB GPU. We run a hyper-parameter search for each of our models and we select the best model based on the dev micro-F1 score. Search values and final hyper-parameters can be found in Appendix A.

6. Results

We present RE results on FRACAS with $mREBEL_{32}$ in Table 6. When comparing overall F1 scores between the two evaluation modes, we first notice that the biggest performance discrepancy seems to originate from the inability of the model to correctly identify the entity tags for (QI) relations (62.07 F1 with $mREBEL_{st}$ versus 66.21 F1 with $mREBEL_{bo}$). This is not surprising, as discriminating between indirect and mixed quotations is quite a difficult task, even for human annotators. In a similar vein, we obtain as expected close performances between the two evaluation modes for (I) relations, given that our dataset models cues with a single entity type (<cue>). This is supported by the fact that (I) relations score is significantly higher than (QI) relations on strict mode (62.07 F1 versus 69.09 F1), although they both have a "Quote" component as the object of the relation, and that this "Quote" component is almost always linked to both a (I) relation and (QI) relation.

In the case of (R) relations, however, both modes surprisingly achieve the same score, which suggests that assigning the correct entity tag to speakers in a (R) relation might be a trivial task for the model, or that extracting co-reference entities is still a difficult task, even with accurate labelling. We presume that the difficulty of (R) RE for FRACAS is strongly related to the repetitive mentioning of speakers in news articles. Even when the right referent is close to the speaker (as explained in 5.1, we made sure referents were found at most two paragraphs away in the text), the appearance of other speaker mentions around the relation may be particularly confusing for the model, especially speakers of the same gender. An in-depth error analysis of $mREBEL_{bo}$ on FRACAS will be conducted to confirm this theory.

7. Conclusion

In this article, we present FRACAS: a new manually-annotated and freely-available resource for quotation extraction and source attribution for French. We use original texts from the Reuters-21578, Distribution 1.0 corpus distributed by NIST that we augment with fine-grained annotations of different types of quotes attributed to their sources and their cues. We adopt an extensive definition of quotation and include Direct, Indirect and Mixed quotes. We obtain a total of 10,965 annotated attribution relations spanning over 1,676 texts. Our annotation process involves 8 anno-

tators and yields satisfying inter-annotator agreement results, which also underlines the complexity of this phenomenon. We use this new language resource to train a state-of-the-art relation extraction model and obtain 67.76 F1 score on the task. These results demonstrate both the usability of generation models for complex relation extraction tasks, especially ones that aim to extract long entities such as quotes, and their usefulness in helping identify the linguistic challenges of this task. Future work on this dataset will focus on extending this resource and using it to better understand the semantic relations between quotes in a text. We then plan to use systems trained on this data to measure gender imbalance in quotations in French media in the context of the *GenderedNew* project (Richard et al., 2022).

8. Corpus data and code

Please visit our Zenodo repository to find the instructions to request the data: https://zenodo.org/record/8353229. The code for our experiments can be found at https://github.com/getalp/fracas-rel-extraction.

9. Limitations and Future Work

As noted in section 3, FRACAS is a relatively small dataset, since it was manually annotated. The data is also unbalanced in terms of class distribution, and highly unbalanced in terms of gender distribution of quotes, which can result into our systems being biased towards disproportionally predicting certain types of quotes, or performing better on predicting quotes by men. We do not investigate this possible bias in this paper, but further work will take into account these limitations. Further experiments on this task could benefit from augmenting this dataset using other freely available data and automatic annotation tools such as the one provided by Cabot et al. (2023). Other work could also evaluate systems tailored for low-resource training on this dataset. Finally, with regard to our proposed experiments, we only describe raw evaluation results due to time constraints, but an in-depth error analysis of named entities is in progress. We however discuss some of the shortcomings of our systems and possible leads for improving RE performance on this dataset.

10. Acknowledgements

We thanks all the annotators who participated in the study and Gilles Bastin. This work is funded by the Université Grenoble Alpes (IRGA ANR-15-IDEX-02) and has been partly supported by MIAI@Grenoble-Alpes (ANR-19-P3IA-0003).

11. Bibliographical References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Mariana S. C. Almeida, Miguel B. Almeida, and André F. T. Martins. 2014. A joint model for quotation attribution and coreference resolution. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 39–48, Gothenburg, Sweden. Association for Computational Linguistics.

Annelen Brunner. 2013. Automatic recognition of speech, thought, and writing representation in German narrative texts. *Literary and Linguistic Computing*, 28(4):563–575.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. Rebel: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381.

Pere-Lluís Huguet Cabot, Simone Tedeschi, Axel-Cyrille Ngonga Ngomo, and Roberto Navigli. 2023. RED^{fm}: a filtered and multilingual relation extraction dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4326–4343, Toronto, Canada. Association for Computational Linguistics.

Eric De la Clergerie, Benoît Sagot, Rosa Stern, Pascal Denis, Gaëlle Recourcé, and Victor Mignot. 2009. Extracting and visualizing quotations from news wires. In *Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2009. Lecture Notes in Computer Science*, Berlin, Heidelberg. Springer.

William Falcon and The PyTorch Lightning team. 2019. PyTorch Lightning.

Emmanuel Giguet and Nadine Lucas. 2004. La détection automatique des citations et des locuteurs dans les textes informatifs. In *Le discours rapporté dans tous ses états: Question de frontières*, pages 410–418. l'Harmattan.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023. DREEAM: Guiding attention with evidence for improving document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1971–1983, Dubrovnik, Croatia. Association for Computational Linguistics.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

Ghassan Mourad. 2000. Présentation de connaissances linguistiques pour le repérage et l'extraction de citations. In *TALN, 7e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles*, pages 495–501, Lausanne, Suisse. ATALA.

Ghassan Mourad and Jean-Pierre Desclés. 2002. Citation textuelle : identification automatique par exploration contextuelle. *Faits de Langues*, (19):13.

Mats Nylund. 2003. Quoting in front-page journalism: Illustrating, evaluating and confirming the news. *Media, Culture & Society*, 25(6):844–851.

Tim O'Keefe, Kellie Webster, James R. Curran, and Irena Koprinska. 2013. Examining the impact of coreference resolution on quote attribution. In *Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013)*, pages 43–52, Brisbane, Australia.

Sean Papay and Sebastian Padó. 2019. Quotation detection and classification with a corpus-agnostic model. In *Proceedings - Natural Language Processing in a Deep Learning World*, pages 888–894, Varna, Bulgaria. Incoma Ltd., Shoumen, Bulgaria.

Silvia Pareti. 2015. *Attribution: A Computational Approach*. Thesis, The University of Edinburgh, Edinburgh, UK.

Silvia Pareti, Tim O'Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska. 2013. Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 989–999, Seattle, Washington, USA. Association for Computational Linguistics.

Fabien Poulard. 2008. Analyse quantitative et qualitative de citations extraites d'un corpus journalistique. In *Actes de la 15ème conférence sur le Traitement Automatique des Langues Naturelles. Rencontres jeunes Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 100–109, Avignon, France. ATALA.

Fabien Poulard, Thierry Waszak, Nicolas Hernandez, and Patrice Bellot. 2008. Repérage de citations, classification des styles de discours rapporté et identification des constituants citationnels en écrits journalistiques. In *Actes de la 15ème conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 141–150, Avignon, France. ATALA.

Bruno Pouliquen, Ralf Steinberger, and Clive Best. 2007. Automatic detection of quotations in multilingual news. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP*, Borovets, Bulgaria.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics.

Ange Richard, Gilles Bastin, and François Portet. 2022. GenderedNews: Une approche computationnelle des écarts de représentation des genres dans la presse française. ArXiv:2202.05682 [cs].

Benoît Sagot, Laurence Danlos, and Rosa Stern. 2010. A lexicon of french quotation verbs for automatic quotation extraction. In *Proceedings of the 7th international conference on Language Resources and Evaluation - LREC 2010*, Valletta, Malta. ELRA.

Andrew Salway, Paul Meurer, and Knut Hofland. 2017. Quote extraction and attribution from Norwegian newspapers. In *Proceedings of the 21st Nordic Conference of Computational Linguistics*, pages 293–297, Gothenburg, Sweden. Linköping University Electronic Press.

Luís Sarmento and Sérgio Nunes. 2009. Automatic extraction of quotes and topics from news

feeds. In *4th Doctoral Symposium on Informatics Engineering*, Porto, Portugal.

Christian Scheible, Roman Klinger, and Sebastian Padó. 2016. Model architectures for quotation detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1736–1745, Berlin, Germany. Association for Computational Linguistics.

Valentin-Gabriel Soumah, Prashanth Rao, Philipp Eibl, and Maite Taboada. 2023. Radar de parité: An NLP system to measure gender representation in French news stories. In *Proceedings - The 36th Canadian Conference on Artificial Intelligence*, Montréal, Canada.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France. Association for Computational Linguistics.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681, Dublin, Ireland. Association for Computational Linguistics.

Jingxuan Tu, Marc Verhagen, Brent Cochran, and James Pustejovsky. 2021. Exploration and discovery of the COVID-19 literature through semantic visualization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 76–87, Online. Association for Computational Linguistics.

Ngoc Duyen Tanja Tu, Markus Krug, and Annelen Brunner. 2019. Automatic recognition of direct speech without quotation marks. a rule- based approach. In *Proceedings of Digital Humanities: multimedial & multimodal*, pages 87–89, Frankfurt am Main.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. ArXiv:1909.03546.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2022. DeepStruct: Pretraining of language models for structure prediction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 803–823, Dublin, Ireland. Association for Computational Linguistics.

Jue Wang and Wei Lu. 2020. Two are better than one: Joint entity and relation extraction with table-sequence encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1706–1721, Online. Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4904–4917, Dublin, Ireland. Association for Computational Linguistics.

Ningyu Zhang, Xin Xu, Liankuan Tao, Haiyang Yu, Hongbin Ye, Shuofei Qiao, Xin Xie, Xiang Chen, Zhoubo Li, Lei Li, Xiaozhuan Liang, Yunzhi Yao, Shumin Deng, Peng Wang, Wen Zhang, Zhenru Zhang, Chuanqi Tan, Qiang Chen, Feiyu Xiong, Fei Huang, Guozhou Zheng, and Huajun Chen. 2023. DeepKE: A deep learning based knowledge extraction toolkit for knowledge base population. ArXiv:2201.03335.

Muitze Zulaika, Xabier Saralegi, and Iñaki San Vicente. 2022. Measuring presence of women and men as information sources in news. In *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 126–134, Gyeongju, Republic of Korea. Association for Computational Linguistics.

12. Language Resource References

Revanth Gangi Reddy, Heba Elfardy, Hou Pong Chan, Kevin Small, and Heng Ji. 2023. SumREN: Summarizing reported speech about events in news. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):12808–12817.

Edward Newell, Drew Margolin, and Derek Ruths. 2018. An attribution relations corpus for political news. In *Proceedings of the Eleventh International Conference on Language Resources and*

Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

NIST. 2005. RCV2 Reuters corpus, National Institute of Standards and Technology, Release date 2005-05-31, Format version 1, https://trec.nist.gov/data/reuters/reuters.html.

Sean Papay and Sebastian Padó. 2020. RiQuA: A corpus of rich quotation annotation for english literary text. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 835–841, Marseille, France. European Language Resources Association.

Silvia Pareti. 2012. A database of attribution relations. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC12)*, pages 3213–3217, Istanbul, Turkey. European Language Resources Association (ELRA).

Timoté Vaucher, Andreas Spitz, Michele Catasta, and Robert West. 2021. Quotebank: A corpus of quotations from a decade of news. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, Virtual Event, Israel.

A. Appendix

A.1. Hyper-parameter Tuning

For all of our models, we tune the learning rate using the Multi-objective Tree-Structured Parzen Estimator implementation from Optuna (Akiba et al., 2019) and run a grid search over warm up steps, batch size, and weight decay. Searched values and best hyper-parameters for each model are shown in the table below.

Grid search values (lr=5e-05)		$mREBEL_{bo}$	$mREBEL_{st}$
Warmup steps	0, 100, 200	200	200
Training batch size	4, 8, 16, 32	32	32
Weight decay	0.0, 0.01, 0.1	0.01	0.01

Table 7: Values tested during hyper-parameter tuning for all models. Best values for the final models are shown in the last two columns.