ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

EUROPEAN BROADCASTING UNION

MASTER THESIS

Towards Chapterisation of Podcasts

Detection of Host and Structuring Questions in Radio Transcripts

MARIN PIGUET

SUPERVISORS: Maud Ehrmann (DHLAB, EPFL), Emanuela Boros (DHLAB, EPFL), Alexandre Rouxel (EBU)

28th March 2024

Abstract - EN

This Master thesis investigates the application of Bidirectional Encoder Representations from Transformers (BERT) on podcast to identify the host and detect structuring questions within each episode. This research is conducted on an annotated dataset of automatic transcriptions of 38 French podcasts of Radio France and 37 TV shows in English of France 24. A variety of BERT models, with different language orientations, are tested and compared on two classifying tasks: the detection of host sentences and the classification of structuring questions. The latter is firstly performed as a three label classification task. Secondly, a reduction to a binary classifier is proposed, with two new configurations.
Initially, BERT models are fine-tuned separately on French and English datasets, as well as on the joint dataset. In a second time, a multilingual approach is implemented with an automatic translation of the original dataset into a total of twenty languages. The translated datasets are used for multilingual fine-tuning and German is included as an evaluation language.
BERT models demonstrate adequate performance in host detection to pinpoint within the list of the speakers the actual host of the show, as well as a proposed comparison rule-based method. For structuring question detection, the three label classifier appears too subtle, at least regarding the size of fine-tuning data. One binary classification configuration yields promising results. The multilingual experiment shows that automatic translation has potential as a source of fine-tuning data and highlight the need for original testing data in these languages.

Abstract - FR

Ce mémoire applique des modèles « Bidirectional Encoder Representations from Transformers » (BERT) sur des podcasts afin d'identifier l'animateur.rice et de détecter des questions structurantes. Ce projet de Master est conduit sur un corpus bilingue de 75 émissions (38 émissions de radio en français et 37 émissions télévisuelles en anglais), provenant respectivement de Radio France et de France 24, tous transcrits automatiquement. Plusieurs modèles de BERT sont testés et comparés sur deux tâches : la détection des phrases de l'animateur.rice, ainsi que la détection de questions structurantes. La détection des questions structurantes est d'abord abordée comme une classification à trois étiquettes, puis simplifiée en deux configurations à deux étiquettes.
L'ajustement fin des modèles BERT est d'abord effectué sur le français et l'anglais séparément, ainsi que sur les deux jeux de données réunis. Dans un deuxième temps, une extension multilingue est réalisée en traduisant automatiquement le corpus initial dans vingt langues différentes. Ces traductions sont utilisées pour ajuster finement les modèles sur toutes ces langues et pour évaluer les performances sur l'allemand.
Les modèles BERT obtiennent des résultats satisfaisants dans l'identification des phrases de l'animateur.rice pour localiser, parmi la liste des locuteurs, la personne en question. Pour la détection des questions structurantes, la tâche à trois étiquettes semble trop subtile, en tout cas par rapport à la quantité de données d'ajustement à disposition. En revanche, l'une des deux configurations obtient des résultats prometteurs. L'extension multilingue démontre l'utilité des traductions automatiques pour l'ajustement fin des modèles et met en évidence l'intérêt potentiel de disposer de données de test originales dans ces langues.

“ *Pour l’enfant, amoureux d’audio à la demande,*
L’univers est égal à son vaste appétit. ”

Charles Baudelaire
(ou presque)

REMERCIEMENTS

Avant de prendre l'antenne, j'aimerais adresser mes remerciements les plus sincères et les plus chaleureux à toutes les personnes qui, de près ou de loin, m'ont assisté dans la réalisation, l'écriture et le montage de ce mémoire.

Mes pensées vont en premier lieu vers celles et ceux qui m'ont supervisé au sein de l'UER comme de l'EPFL. Au Grand-Saconnex, Alexandre Rouxel, pour sa bienveillance, l'intelligence de ses conseils et pour m'avoir ouvert la porte de cette précieuse institution. Egalement Pierre Fouche pour sa gentillesse et sa disponibilité permanente à l'assistance technique. A Ecublens, Maud Ehrmann, pour sa supervision académique et ses relectures rigoureuses. Aussi, Emanuela Boros pour sa compétence et son assistance active. Puis, en duplex de Paris, toute l'équipe de l'unité recherche et développement de Radio France, à commencer par Ivan Thomas, dont la bonne humeur permanente, la curiosité inébranlable et la gratitude chaleureuse furent déterminantes. Je n'y oublie pas non plus Jon Stark, Allaoua Benchikh et Matthieu Beauval.

Au-delà de ce semestre de projet de Master, je ne saurais trouver les mots pour exprimer ma gratitude envers toutes celles et ceux qui m'ont soutenu tout au long de ce parcours sinueux que furent ces neuf années à l'EPFL. Tout aurait été bien différent si je n'y avais pas rencontré toutes ces personnes. Marko, Hannah, Lola, Pierre et tant d'autres...

Bien évidemment, je serais honteusement incomplet si je ne mentionnais pas le soutien apporté par tout mon entourage extra-académique : les chœurs, les orchestres, les professeurs de musique, les collègues de la Radio Télévision Suisse, et naturellement toute ma famille, mes parents en tête.

AND TWELVE POINTS GOES TO ...

Finalement, il y a celui sans qui aucune de ces lignes n'aurait vu le jour. Daniel Rausis, qui m'a pris sous son aile il y a quatorze ans déjà, dans cette drôle émission d'Espace 2, *A vous de jouer*. Chaque instant passé à ses côtés n'a fait qu'aiguiser mon regard sur le monde et accroître le plaisir de l'expression au micro et ma passion pour le monde de la radio. Au moment de remettre ce mémoire, la veille même de son départ à la retraite, je tiens à lui dédier ce travail et à lui souhaiter mes vœux les plus fervents pour l'avenir.

Contents

CHAPTER 1

INTRODUCTION

Welcome to this Master's thesis in Digital Humanities. Today, live from EPFL: how can we structure audio content, using natural language processing methods, and make podcasts more easily explorable?
Thanks for joining us!

1.1 IT ALWAYS STARTS WITH A JINGLE

Since decades, audio broadcasting has been a central element in human consumption habits in terms of entertainment, information or education. With lower need of attention than video content, audio can accompany auditors in many of their activities. This type of media was in the first place the privilege of radios and in particular of public service media which for some, like the Radio Télévision Suisse in 2022 or the Czech Radio (Český rozhlas) in 2023, are celebrating their 100th anniversary at that time. Nowadays, these traditional broadcasters propose their programs by many other means than the FM band, for example on the Internet, where the number of available platforms has strongly increased in recent years and also allowed the development of a huge variety of independent or amateur productions. The teeming variety of shows, in terms of content, format or support is now gathered together under a single name: podcasts. Today, they define any type of audio or video content that is available online for streaming or downloading.

Paradoxically, sharing audio content and making it explorable outside a radio stream remains challenging. It is in this context that Radio France, the French public service radio, encouraged by its televisual cousin France 24, turned to the European Broadcasting Union (EBU), with the need of a data science approach for conceiving an automatic chapterisation method. Podcasts being easily an hour long, it can be valuable for any user, both in terms of daily consumption or information retrieval, to have prior knowledge on the content with more than a short summary. This information can be the people talking in the podcast, the topics covered, also possibly within digressions, or the tonality of the discussion (for example if a subject is presented in a positive or negative manner). With the intuition that each podcast must have a pre-established construction adapted to the whim of conversation, it should be possible to extract freestanding parts, or chapters, and help to give a coherent and detailed description of the whole podcast.

1.1.1 PARTNERS OF THE PROJECT

The EBU is not a radio producer strictly speaking. It is an international organisation founded in 1950 by the first European broadcasters and currently has more than one hundred European public radio and

television stations as members. The EBU represents their broadcasting rights, manages the Eurovision and Euroradio networks and takes the initiative of the organisation of many events, from internal technical conferences to the well-known Eurovision Song Contest. The EBU has a close and privileged relationship with its Members, which allows a trusting collaboration with them also in a research perspective.

Radio France is the French public service radio. It was founded in 1975, but is the heir of previous public institutions responsible for broadcasting since 1939. Radio France is today at the head of many radio channels, the most important being France Inter, France Culture, France Info, France Musique and the forty-four local versions of France Bleu. These channels offer a huge variety of podcasts, like interviews, magazines, debates, documentaries, morning shows, concerts retransmissions, etc. The activity of Radio France is not limited to radio production, but also includes the organisation of many musical events and the management of two symphonic orchestras and two choirs.

France 24 is a French public service international-minded TV channel created in 2006. It is a member of France Médias Monde and therefore also of the EBU. It is broadcast world-wide in four languages (French, English, Spanish and Arabic) and covers continuously international news with regular flashes interspersed with debates, documentaries or magazines.

1.2 COURSE OF A PODCAST AND GLOSSARY

Even if the content and the form of a podcast can vary a lot between the types of show, this section defines the typical situation that is handled in this thesis and proposes a small glossary on the main terms of the radio world that can be used in an ambiguous way sometimes.

1.2.1 SETTING OF A RADIO SHOW

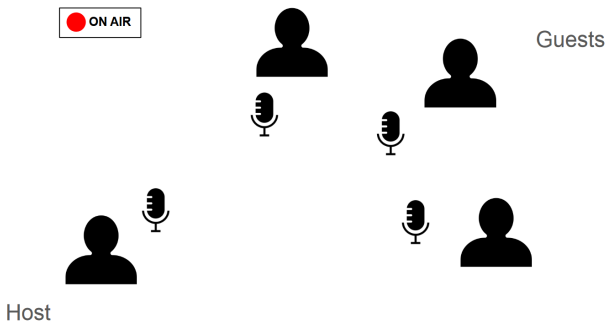


FIGURE 1.1
Typical setting of a show.

In Figure 1.1, the most common situation is represented. In the podcasts relating to the present work, a show is typically conducted by a host who is also, in general, the producer of the podcast. The podcast is defined by a mandate that outlines the global topics to be covered in the various episodes and provides guidelines on how these subjects should be addressed, in terms of target audience, type of elements (columns, documentaries, interviews, etc.) or profiles of expert guests. During the podcasts of this project, the host converses with the guests, a small group of experts, usually between one and four. The host intervenes on a regular basis by asking some prepared questions, but also ask for reactions on elements that appear in the course of conversation and presents the elements that punctuate the show.

1.2.2 GLOSSARY

Anchor A radio employee who presents a radio program in a more formal manner than a *host*. Its role is to deliver a text in a more authoritative manner and without any form of interactivity.

Guest An expert of any field who is invited to talk about it in a *show*.

Host A radio employee who is the presenter of a *show*, and generally the producer of the *podcast*. Its role is to present the topics, the music or any kind of audio element, to introduce the *guests* and to interview them. The structure of the *show* and its organisation are under its responsibility.

Podcast A radio production that consists in a series or collection of episodes, broadcast on a regular basis.

Show An occurrence of a *podcast*, often on a daily or weekly basis.

1.3 OBJECTIVES

The goal of this Master thesis is to develop an automatic chapterisation algorithm, based on the structuring questions that the host ask during the show.

The standard online representation of the podcasts nowadays can be seen in Figure 1.2. It can be seen that at that point, the prior knowledge that one can have on the podcast before listening to it is the title of the podcast, the secondary title of the particular show, the experts that are invited and their title, a small introduction, which is in that case the written transcription of the real intro of the podcast that was probably copied from the notes of the host, and finally a small and very succinct summary of the content, with one specific description of a report presented in the second part of the show. Not visible in the Figure, in this particular podcast, Radio France recommands a list of references that are related to the subject.

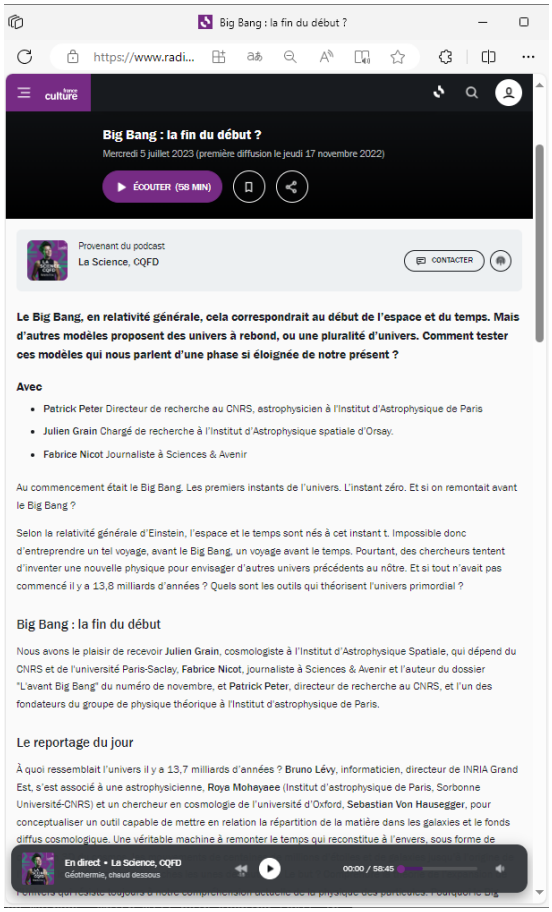


FIGURE 1.2
Current "representation" of the podcast on the website of Radio France.

One can admit that this information is not exhaustive, and in particular, that none is connected to some specific timestamps in the show, which could be valuable for a faster, partial or non-linear listening. In addition to that, it asks for human labour to shape it properly. Therefore, automatically dividing the show into chapters is becoming necessary as a means of providing a new listening experience for the audience.

With the emergence of language models in recent years, both text analysis (on a broad point of view) and speech-to-text algorithms have increased their performances. It is therefore a legitimate question to think about the assistance such tools can provide for showcasing podcasts and make them more easily explorable and researchable. In this thesis, podcasts are analysed from *transcriptions*, but the performances of the automatisation of the transcription task is not evaluated in this work, considering it as sufficiently good to work on the transcript as analysis material. The same applies for the recognition of the different speakers in the show, which is called *diarisation*.

This work is confronted to many challenges, in particular the industrial context of this research and the specific needs it implies. For the EBU, the objective is to develop a prototype platform providing AI tools for its Members for several aspects of the audiovisual sector, including a podcast analyser called

Meta Radio[1] that should include some of the products of the present research. In consequence, it is necessary to be able to cope with the drastic scope of languages of the EBU Members, covering every living languages with Indo-European or Uralic origins, and even some Turkic and Semitic languages. The research starts with the French content of Radio France and the English one of France 24 and later on proposes a multilingual extension method.

1.4 TASKS

The objective of this Master thesis is achieved by dividing the process of chapterisation into two steps and, therefore, developing two distinct classifying tasks: the identification of the host and the identification of structuring questions. These questions are considered as key segmenting points and change of chapter indicators.

The first task, i.e. host detection, is conducted using two approaches. The first approach involves using diarisation and a rule-based method to identify the host. The second approach involves the transcription and fine-tuning Bidirectional Encoder Representations from Transformers (BERT) models. To be more precise, the relevant questions for this problematic are:

- Which BERT models identify the best the sentences pronounced by the host?
- Does it perform better in French or in English?
- What are the performances of BERT models in comparison with a rule-based method?
- Could the rule-based method and BERT models be complementary?

To properly answer these questions, for both methods, each sentence is predicted whether it has been pronounced by the host or not, and this is compared with a ground-truth.

The second task is the identification of important questions asked by the host, that give a clear structure to the show. This is performed with fine-tuning BERT models. The questions are:

- Which BERT models identify the best the structuring questions?
- Does it perform better in French or in English?
- How can the detection of the host be complementary to this task, if it is performed beforehand?
- Finally, in the perpective of the multilingual extension, can automatic translation tools be a way of evaluating the transferability of the models to other languages? Can translation also be a way of fine-tuning the models on other languages?

These questions are also answered by comparing the prediction for each sentence with a ground-truth annotation.

Section 2 presents related work on text segmentation, host detection and BERT principles. Section 3 presents the dataset of podcasts: how it was transcribed and annotated, and the statistics of its content. Section 4 presents and discusses the experiments made on host detection and identification of structuring questions, while Section 5 explore its transfer learning to other languages. In the end, Section 6 opens the methodology of structuring question detection to encompass more specific features of the podcast sphere, and Section 7 discusses the global results and the potential enhancements and continuations of this work.

[1]The tool was presented at the Data Technology Seminar on March 13, 2024, by Alexandre Rouxel (for generic purposes), Ivan Thomas (for industrial needs) and Marin Piguet (for the implementation of the work described in this thesis): `https://tech.ebu.ch/publications/presentations/dts2024/the_ebu_ai_hub_radio_enrichment`

Figure 1.3

Augmented podcast representation featuring a direct link between the transcription and the audio player. Chapters appear on the left and could, for example, start at the most structuring questions of the podcast. The chapters can be automatically titled and summarised, as shown in this example, using a generative large language model.

CHAPTER 2

RELATED WORK

Now that our audience is familiar with the context and the goals of the project, what can you say about the existing work on the subject? Is it completely original? Or can we find publications with similar approaches or objectives?

The raw speech-to-text transcriptions can be enriched in many manners. Undoubtedly, since the podcasts were chosen to be processed from that textual perspective, it is relevant to take a look around and get inspired by work that has been done on written data. Especially as existing audio segmentation algorithms are also based on textual transcriptions, as it is shown in this section.

The first straightforward idea for such long texts is to partition them into coherent and freestanding chapters. However, even this objective can mean many things. The most standard approach is segmenting the texts from the perspective of the topics they talk about. One well-known task is called Text Tiling and consists in detecting ruptures in a text, where a change of topic occurs, thanks to some particular features or embeddings. Different research on the subject are shown in Section 2.2. However, none is based on podcast transcriptions. Furthermore, none uses any particular element from the form or document structure of their respective datasets, as the host is used in this study, and they are strictly centred on the content. Nevertheless, the method proposed in this work is related to Text Tiling in that it also seeks to identify break points in the text, with the difference that when Text Tiling aims for the disparities or maximum distance on both sides of these points, the present work targets sentences that are the initiators of these breaking points. With this in mind, the detection of the host is obviously a first crucial matter of focus of this work, therefore some existing methods and work are presented in Section 2.3.

The two classification tasks developed here are using, as mentioned, Bidirectional Encoder Representations from Transformers (BERT). Therefore, this section starts with their presentation.

2.1 BERT

BERT is a language representation model that was introduced in 2018 by Devlin et al. On its release, it was presented as a model that achieved state-of-the-art performances on a wide scope of Natural Language Processing (NLP) tasks and that can be easily adapted to specific purposes (Devlin et al. 2018). Since the beginning, there are in fact different versions of the model, but they all share some fundamental designs, such as being neural networks.

BERT has a multi-layer architecture that is divided in two main frameworks: the pre-training and the fine-tuning. This division makes it convenient to use, since all the expensive cost is concentrated on the pre-training layers and that only one additional layer is required in the fine-tuning step, which takes reasonable time and resources. In more details, BERT models are pre-trained on two tasks with unlabelled data. These tasks are called *masked language model* (MLM) and *next sentence prediction* (NSP). MLM consists in masking some token of a sentence and then predicting the token from the context, when NSP consists in determining if, in a pair of sentences, the second sentence is the continuation of the first. From then on, the fine-tuning step starts from this initial parametrisation to specialise the model on labelled data for a new specific task. The result is called the *final downstream architecture* and is rather close to the original pre-trained architecture.

The initial model was built in two sizes: $\text{BERT}_{\text{BASE}}$ that has 12 layers, for a total of 110M of parameters, and $\text{BERT}_{\text{LARGE}}$ that has 24 layers, for a total of 340M parameters. They were both trained on two datasets: BookCorpus[1] and English Wikipedia[2]. The two models are themselves available in two forms: one that is sensitive to the case of the text (`cased`), recognizing the difference between capitalized and non-capitalized words, and another that is not case-sensitive (`uncased`), treating all words as if they were in lowercase.

From these initial models, `bert-base-uncased`, `bert-base-cased` and `bert-large-uncased` are used in the present work.

Model	Language	Pre-training data	#parameters
`bert-base-uncased`	English	BookCorpus & English Wikipedia	110M
`bert-base-cased`			110M
`bert-large-uncased`			340M
`camembert-base`	French	OSCAR	110M
`camembert-large`		CCNet	335M
`bert-base-german-cased`	German	German Wikipedia, OpenLegalData dump & new articles	110M
`xlm-roberta-base`	Multilingual	2.5TB of filtered CommonCrawl data	279M
`xlm-roberta-large`			561M
`bert-base-multilingual-cased`		Wikipedia	179M

TABLE 2.1
BERT models employed in the current study.

2.1.1 BERT ECOSYSTEM

BERT was thereafter quickly adapted to other languages than English, with the same architecture trained on new data. For this study, the base and large versions of `camembert`[3] are tested and used for handling the French data. The base model is pre-trained on the OSCAR[4] dataset and the large on CCNet[5]. In the meantime, some multilingual models undergo evaluation to handle the two initial languages and, later on, a wider scope of them. The multilingual models are the base and large version of `xlm-roberta` and `bert-base-multilingual-cased`. The latter is a multilingual pre-training on Wikipedia data of the initial BERT model, when the first is the multilingual variant of `roberta`, that was itself a later and

[1] `https://yknzhu.wixsite.com/mbweb`
[2] `https://en.wikipedia.org/wiki/English_Wikipedia`
[3] `https://camembert-model.fr/`
[4] `https://oscar-project.org`
[5] `https://github.com/facebookresearch/cc_net`

more robust implementation of BERT, pre-trained on data filtered from CommonCrawl[6]. In the end, the German-based model `bert-base-german-cased` is also used. It is trained on various data, such as German Wikipedia[7], OpenLegalData dump[8] and new articles. A summary of all the models that are included in this study can be seen in Table 2.1.

2.2 TEXT TILING

Text Tiling was first introduced by Hearst, who defined it as the "partition [of] texts into contiguous, non-overlapping subtopic segments" (Hearst 1997). Hearst concedes that the notion of *subtopic*, and even of *topic* itself, although intuitive at first sight, is complex and subject to many pages of discussions or definitions. Therefore, he suggests investigating for key points of *changes or shift of topic* rather than trying to identify the topic related to each part of the document that is to be segmented. His method for this investigation is purely based on lexical and similarity score comparison and is rather outdated, but the search for these breaking points is still a matter of study nowadays. If in this work, these key transitions are sought among the sentences of the host, alternative approaches from recent years can be considered.

Yoong, Fan and Leu addressed Text Tiling with BERT-based models with three different methods (Yoong, Fan and Leu 2021). The first one is BERT-NSP that reuses the primary *next sentence prediction* task of BERT to determine whether a given sentence is the following of another from a topical viewpoint. It is run over all adjacent sentences in a document to determine the probability of having a topic-shift at each sentence. The two others are BERT-SEP and BERT-SEGMENT that both consider the whole document. BERT-SEP at sentence-level representation and BERT-SEGMENT at token-level representation. In both cases, sentences are separated by the `[SEP]` BERT token. After computation of BERT embeddings, BERT-SEP computes the probability of a topic-shift at each $\mathbf{V}_{\texttt{[SEP]}}$, when BERT-SEGMENT computes it for each token of the document and aggregates all score within each sentence to get the probability at the sentences level. In this work, experiments are made on three benchmark datasets made of Wikipedia articles. Both BERT-SEP and BERT-SEGMENT methods yield better results than previous state-of-the-art text segmentation algorithms, but the best result is obtained with one of their variants of BERT-SEP. This can be taken as an incitation to work at sentence level also for podcasts. However, these two first research articles are based on texts that are initially in written form. One should take a look at some work done, if not on podcasts segmentation, at least on audio documents.

Solbiati et al. performs chapterisation on meeting recording transcripts (Solbiati et al. 2021). They were processed in an unsupervised way, from a topic perspective, pushing the work of (Hearst 1997) further by computing similarity with BERT embeddings, rather than with classical term-frequencies. Although this method may be seen as a step closer to podcast segmentation, as it is based on two audio datasets, its fundamental nature remains similar to textual chapterisation. The main distinction lies in the fact that meetings lack the formal and codified structure that podcasts can inherit from hours of conceptualisation, pre-production, and post-production.

2.3 HOST DETECTION FOR CHAPTERISATION

On a broad point of view, what are the qualities of a host? In the sense, what makes it distinguishable? Heiselberg and Have studied the expectation of listeners regarding the host of the shows they listen to (Heiselberg and Have 2023). The study reveals three key qualities: knowledge, storytelling and parasocial relationship. The host has to show a wide curiosity and a familiarity with the topics covered

[6] `https://commoncrawl.org/`
[7] `https://en.wikipedia.org/wiki/German_Wikipedia`
[8] `http://openlegaldata.io/research/2019/02/19/court-decision-dataset.html`

(knowledge) on one hand, and on the other hand must be well-prepared and concise (storytelling). From the perspective of this thesis, these can be characteristics that can be perceived from the texts of the podcasts themselves, and therefore potentially identified by a Large Language Model. The parasocial relationship, however, being more related to its engagement, its mood or personality, may be more conveyed by audio characteristics, and therefore not directly related to the present work. That being established, the discussion of Heiselberg and Have insists on the fact that "the messenger matters", showing, if it was to be demonstrated, the fundamental role that the host plays in a show, and the interest in focusing on its speaking turns for segmentation and chapterisation.

The state of the art of host detection is more dedicated to identifying an anchor rather than a host. The differences between the two roles makes the methods not fully transferable from one situation to another, but the common characteristics of the two and their similar role in structuring and chapterisation makes it interesting to be compared.

Charlet proposes a detection approach of the anchor for TV news shows (Charlet 2010). In this case, the dataset is made of 38 French news programs from different broadcasters. Interestingly, even if working on an audiovisual dataset, her method is purely based on the audio and the anchor detected with speaker clustering method and rule-based criteria. The different audio segments of the shows are clustered with a personalised implementation of a standard algorithm. After clustering, the host is linked to a cluster following a rule-based approach. This method allows performing some reclustering after selection of anchor's cluster, contrary to the present work (see Section 3, where the clustering is left to an independent tool and the parameters not preserved). Charlet chose three criteria to select the cluster associated to the anchor. It is supposed to be the one that, firstly, speaks the most, secondly that has the smallest average interval between its speaking turns and finally with the "purest" cluster. These rules are in exact contradiction with the rule-based approach proposed in this thesis (see Section 4.1.1), since the hosts of the podcasts of this project are never the speakers that speak the most, they ask short questions that require long answers, and therefore have not the smallest intervals between speaking turns, and finally, the parameters of the cluster being not preserved, the "purest" cluster cannot be determined. However, they are aligned with this study's hypothesis that only one host or anchor is leading the show (see Section 3.4.1) and demonstrate that a rule-base approach can be effective in an adapted context.

Yella, Varma and Prahallad used anchor detection to produce summaries of broadcast news from the British Broadcast Corporation (BBC) based on sentences extraction (Yella, Varma and Prahallad 2010). As in (Charlet 2010), the change of speakers are determined with standard statistical techniques and anchor is chosen to be the speaker that speaks the most. But the real interest of (Yella, Varma and Prahallad 2010) is chapterisation. The speaking turn of the anchor are considered as beginning of a new chapter, if they are longer than five seconds. Finally, the initial segments of a determined length of each chapter are concatenated to be used as a summary, that are evaluated to be better than standard NLP methods. The present work could be inspired by the sentence extraction in speaking turns of the host for summarisation, however, length criteria would not be accurate, since the role of host is more interactive than informative, as the role of the anchor is.

2.4 DISCUSSION

The existing research in this section undoubtably contain considerable differences from the specific goals and particular context of this thesis. Nevertheless, certain approaches can provide valuable inspiration for the present work. For example, they all highlight the interest of using transcription of podcasts and analyse them from the textual perspective. Furthermore, some work also suggests that even television shows should be addressed from that perspective. They also reveal that customised strategies adapted to the circumstances of the creation of the podcasts are worth giving some time in design.

CHAPTER 3

PODCAST DATASET

> *I think it is time to get to the heart of the matter. What exactly did you use as material? And how did you process it?*

For this project, a dataset of French and English podcasts is built with the help of members of the research and development unit of Radio France that provided the EBU with a suggestion list of podcasts from their own productions. For the English content, they were authorised by their counterpart of France 24 to provide a set of their shows.

(A) Podcasts in French

(B) Podcasts in English

FIGURE 3.1
Podcasts in French are provided by Radio France and the English by France 24.
The bigger the icon is, the more there are shows of the podcast.

Radio France podcasts come from three of their channels: France Inter, France Culture and France Musique. The two first offer a wide range of weekly or daily fifty minutes long topic-specific shows that constitute the majority of the French dataset. These shows are typically conducted by a host who talks about one main subject with one or more (up to three) guests. Some documentaries or political interviews are also present, as well as short artist interviews extracted from more musical minded radios like France Musique. The shows are conceived for a hearing consumption only, without video version. An overview of the selected French podcasts can be seen in Table 3.1 and Figure 3.1a.

Podcast	Radio	# Shows	Type
La science CQFD	France Culture	12	In-depth scientific discussions
Entendez-vous l'éco		4	Economic and financial discussions
Sens politique		2	Political discussions
Les nuits de Culture		1	Archives, discussions, interviews
Affaires étrangères		1	International politic discussions
L'invité des matins d'été		1	Interviews with cultural or political figures
L'invité des matins		1	Interviews with cultural or political figures
Cultures monde		1	Contemporary challenges in international contexts
Toute une vie		1	Personal and biographical radio portraits
Géographie à la carte		1	Geographical discussions
Soft power		1	Cultural and international discussions
Le cours de l'histoire		1	Historical discussions
La série documentaire		1	Documentary series
L'esprit public		1	Political news debated by engaged intellectuals
On aura tout vu	France Inter	1	Cultural and artistic discussions
Totemic		1	French news with foreign journalists
Grand bien vous fasse		1	Health and wellness discussions
Interception		1	Investigative journalism
En quête de politique		1	Political discussions
Une journée particulière		1	Personal and biographical discussions
Questions politiques		1	Political discussions
Le concert du soir (post-concert interview)	France Musique	1	Interviews with musicians
Les grands entretiens		1	Interviews with cultural figures

TABLE 3.1
Overview of the French podcasts of Radio France that constitute the dataset.

Podcast	TV	# Shows	Type
The debate	France 24	18	Political news debated
Talking Europe		6	European political news and analysis
The 51%		4	Women's perspectives on global issues
The interview		3	Interviews with newsmakers and cultural figures
Science		2	Scientific and technological developments
Tech24		2	Technology news and analysis
Perspective		2	In-depth reporting and analysis of global issues

TABLE 3.2
Overview of the English podcasts of France 24 that constitute the dataset.

France 24 podcasts are for the majority more oriented on politics. They focus on personalities interviews or debates. The political coverage of these shows can lead to a more contradicting role for the host than in the French dataset. However, subject-specific shows are not totally absent. The shows are made to be broadcast on television and therefore some information can be missing in the audio part. From a qualitative perspective, all shows should be intelligible for humans just from the audio. An overview of the selection can be seen in Table 3.2 and in Figure 3.1b.

3.1 TRANSCRIPTION AND DIARISATION PIPELINE

Each audio recording of the shows is transcribed into text with the `whisper-timestamped` library[1] (Louradour 2023), using `large-v2` model of the original OpenAI `whisper`[2] (Radford et al. 2022) and with the hyperparameters:

```
whisper.transcribe(model, audio, language="[language of the podcast]",
    beam_size=5, best_of=5, temperature=(0.0, 0.2, 0.4, 0.6, 0.8, 1.0),
    initial_prompt="C'est curieux chez les marins ce besoin de faire des phrases.",
    vad=True)
```

The hyperparameters `beam_size`, `best_of` and `temperature` are the recommended values to use according to `whisper-timestamped` documentation. The `vad` hyperparameter stands for voice activation detection and prevents `whisper` to transcribe text where there is none. Finally, `whisper` is known to occasionally stop punctuating its transcriptions. According to the forum of the library, these faux pas can be reduced by providing an example punctuated sentence as `initial_prompt`.

In the end, the resulting sentences are split, stopping at punctuation characters. The *start* and *end* timestamps of the sentence are determined with the *start* timestamp of the first token and the *end* timestamp of the last one.

In parallel, speaker diarisation is performed with `pyannote.audio` library[3] (Plaquet and Bredin 2023). The latter clusters the different speakers in the show and returns the timestamps of their speech utterances. This process does not determine the identity of the speakers and rather group their utterances under a random speaker ID. Then, in the podcast transcript, each sentence is attributed to the speaker that the timestamps of the sentence overlap the most.

The final result is exported into one Excel file per show, with columns *Sentence*, *Start*, *End* and *Speaker*. Three additional columns are also created *Animateur.rice*, *Structurante* and *Comment*, dedicated to annotation. An example of one line of a transcript can be seen in Table 3.3.

Sentence	Start	End	Speaker	Animateur.rice	Structurante
Hello, I'm Annette Young and welcome to the 51% show about women reshaping our world.	0,92	25,80	SPEAKER_04	Yes	Fort

TABLE 3.3
Example of podcast transcript. The *Sentence*, *Start*, *End* and *Speaker* columns are filled automatically, when *Animateur.rice* and *Structurante* are annotated by hand. The *Comments* is not presented here, since it was only used by annotator to explain qualitatively its way of thinking.

[1] `https://github.com/linto-ai/whisper-timestamped`
[2] `https://openai.com/research/whisper`
[3] `https://github.com/pyannote/pyannote-audio`

3.2 DATASET STATISTICS

The dataset contains a total of 38 shows in French and 37 in English, for a total length of respectively 33h 16 min and 16h 20 min.

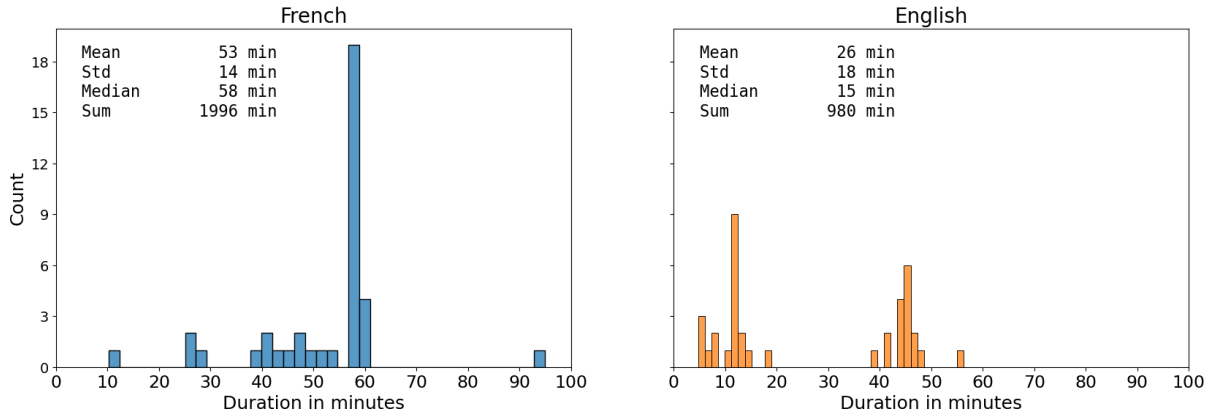


FIGURE 3.2
Duration distribution within French and English shows in the dataset.

The exact distribution of duration over shows in both languages can be seen in Figure 3.2. One can notice a significant difference between the two distributions and the shorter nature of the chosen English podcasts, which also lead to the total length difference between the two languages. This is simply due to the dissimilar types of shows that Radio France and France 24 produce.

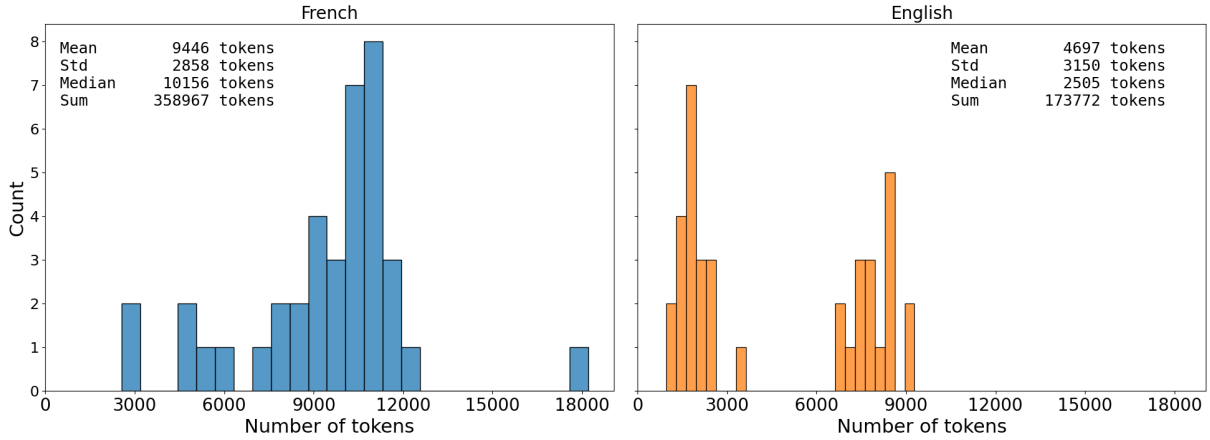


FIGURE 3.3
Token distribution within French and English shows in the dataset.

The French and English datasets contain respectively 358,967 tokens and 173,772 tokens. The token distribution over shows is presented in Figure 3.3. One would notice that if the two distributions, i.e. over durations and tokens, are equivalent in English, the peak around the 60 minutes French podcasts is spread between 9,000 and 12,000 tokens. This might come from the fact that the French dataset is built with different shows that are for most of them one hour long, but can contain a various amount of non-verbal elements (e.g. music, ambiance sounds, etc.).

If the French and English datasets differs from their show duration and tokens repartition within podcasts, the distribution of tokens over the sentences that compose these podcasts are relatively similar. These distributions are shown in Figure 3.4.

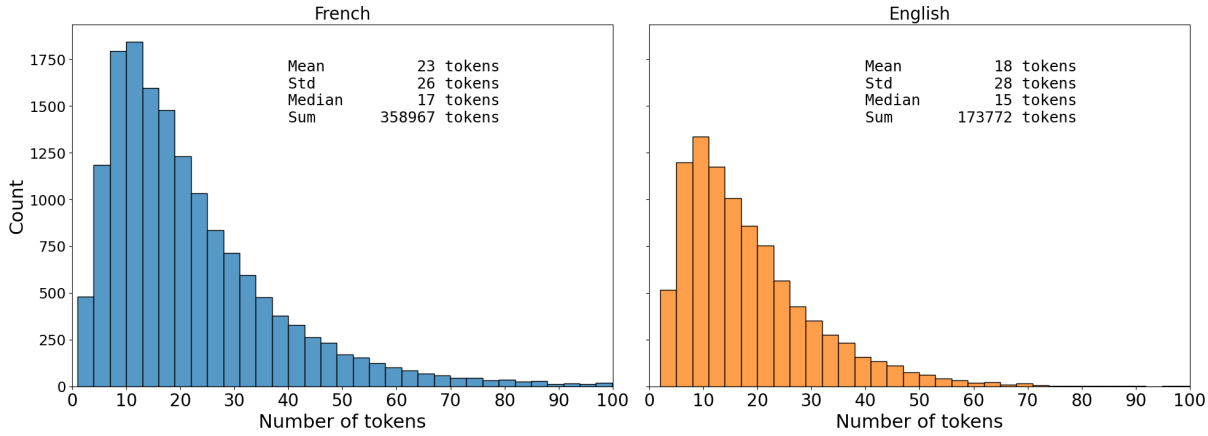


FIGURE 3.4
Sentence length distribution within French and English datasets.

3.3 DATA SPLITS

The dataset is split into three subsets (training, validation and test) to be used during for the experiments. Even if each data point corresponds to a sentence, the 80-10-10% ratio between the three subsets is calculated at the level of the shows, so that all sentences of a show are in the same split. This is done to be able to keep track of continuity between sentences if needed, or to possibly perform a qualitative evaluation of segmentation on a complete show from the test set. Table 3.4 presents the statistics of the datasets, with information about the number of sentences, tokens, durations, etc.

	Data Splits by Language					
	French			**English**		
	Train	**Val**	**Test**	**Train**	**Val**	**Test**
# Shows	30	4	4	29	4	4
# Sent. (All)	12,463	1,683	1,495	7,112	1,095	1,208
# Sent. (Host)	3,202	468	401	2,107	300	357
# Tok. (All)	284,888	37,483	36,596	133,179	20,034	20,559
# Tok. (Host)	59,085	8,246	8,506	34,640	5,093	5,466
Duration	26h 14min	3h 37min	3h 25min	12h 30min	1h 53min	1h 57min
Total	38 Shows, 15,641 Sentences (4,071 Host), 358,967 Tokens (75,837 Host), 33h 16min			37 Shows, 9,415 Sentences (2,764 Host), 173,772 Tokens (45,199 Host), 16h 20min		
Overall Total	75 Shows, 25,056 Sentences (6,835 Host), 532,739 Tokens (121,036 Host), 49h 36min					

TABLE 3.4
Detailed repartition of data within the three subsets by language, including all specified categories and totals.

3.4 ANNOTATION PROCESS AND LABELS

The transcribed shows are then annotated by *The Evaluations and Language resources Distribution Agency*[4], focusing on two elements: the host of the show (*Animateur.rice* column) and the structuring nature of the sentence (*Structurante* column).

3.4.1 HOST OR *Animateur.rice*

This column is annotated **1**, if the speaker that pronounced it is the host of the show. Even if the selection only includes podcasts with a single host, it can happen that a supplementary radio employee appears momentarily in a show. This can typically happen when someone comes for a radio column or presenting a reportage. In this case, the column is annotated **2**, or even **3**, etc. These speakers will be considered as host or not according to each situation. In any other case, the column is annotated **0**.

3.4.2 STRUCTURING OR *Structurante*

The question (in a very broad meaning) of the host is considered as *structurante* if it initiates a change of topic in the discussion. It can take three discrete values: **Non**, **Faible** or **Fort** (meaning respectively **No**, **Weak** and **Strong**).

- Only the sentences of the host are annotated **fort** or **faible**.
- A general question such as *"How are you doing?"* does not structure the discussion, therefore the label **non** is used in that case.
- The label **non** is also used for host introductions, since they do not initiate a topic switch.
- When the host asks a guest to react to something that has been said, the label **faible** is used, since it creates a substructure in the show.
- If the host repeats something that has been said, to ask for precision, the label **faible** is used. This is a typical case where a sentence can be considered as a question, even if no question mark is required at the end of the sentence.

[4] `http://www.elra.info/en/about/elda/`

3.5 HOST STATISTICS

After annotation, some more detailed statistics can be made, in particular what proportion of the sentences are pronounced by the host.

The distributions of host tokens within shows and within sentences can be found respectively in Figure 3.5 and Figure 3.6. For French podcasts, 75,837 tokens are said by the host, which correspond to 21% of them. For English podcasts, these are 45,199 tokens, corresponding to 26% of them.

One can also examine the distribution of the three different labels of *Structurante* class. Figures 3.7 and 3.8 represents the distributions within the whole dataset, respectively the sentences of the host.

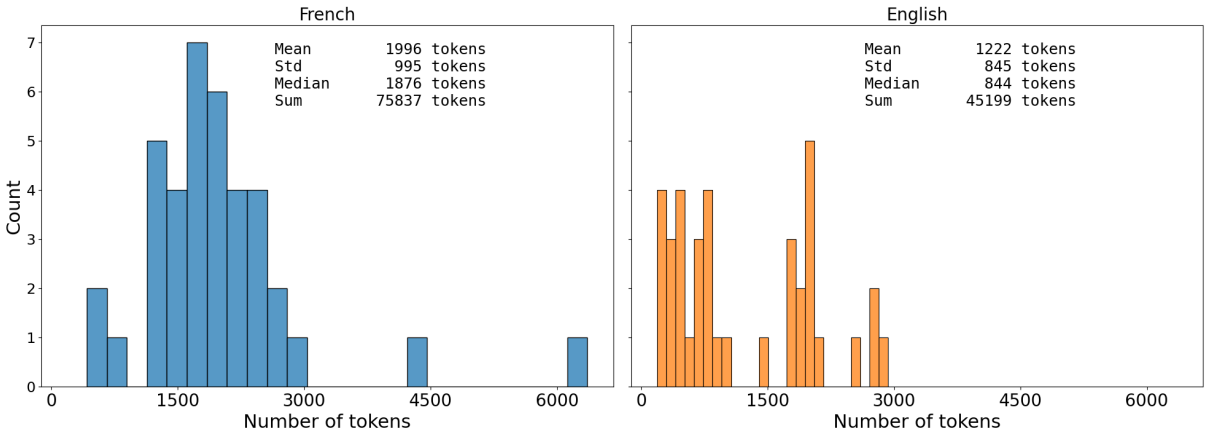


FIGURE 3.5
Host token distribution within French and English shows in the dataset.

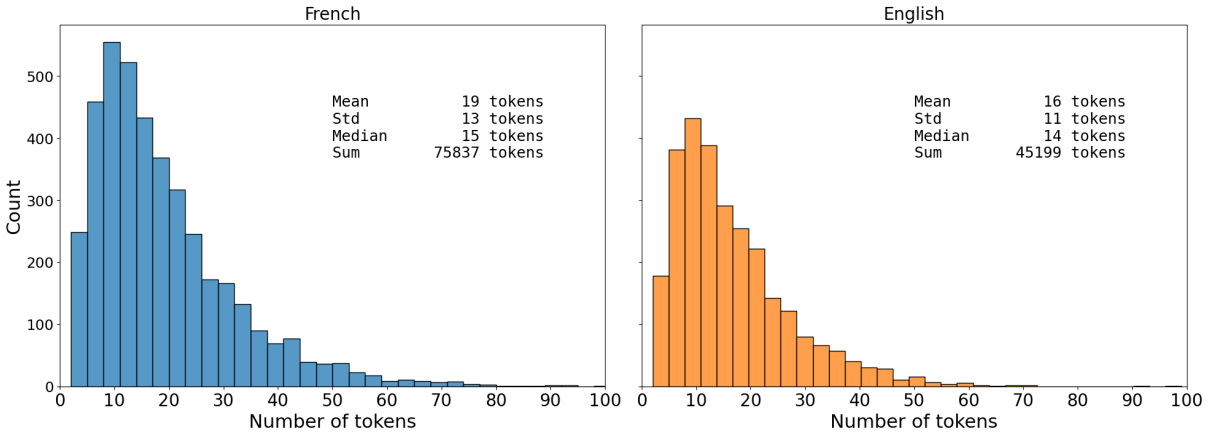


FIGURE 3.6
Host token distribution within French and English sentences in the dataset.

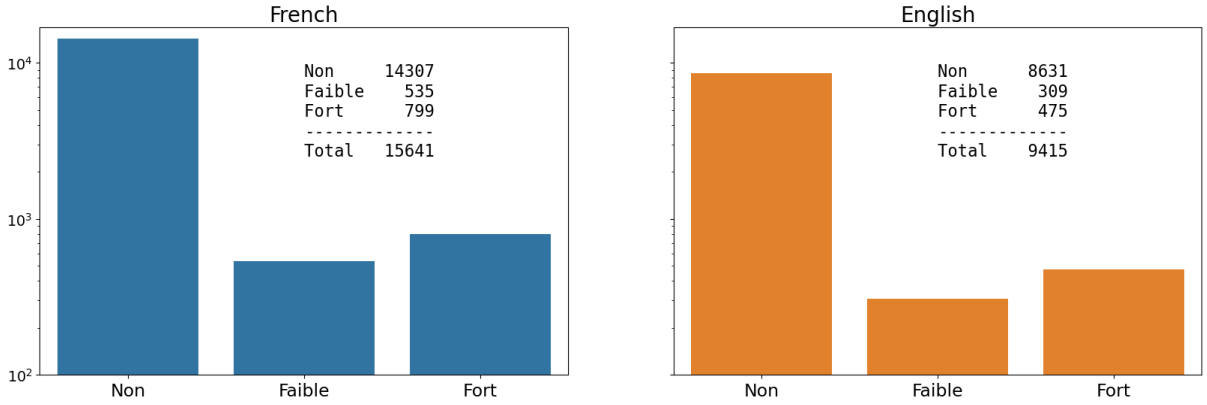


FIGURE 3.7
Class distribution within French and English in the dataset (on log scale).

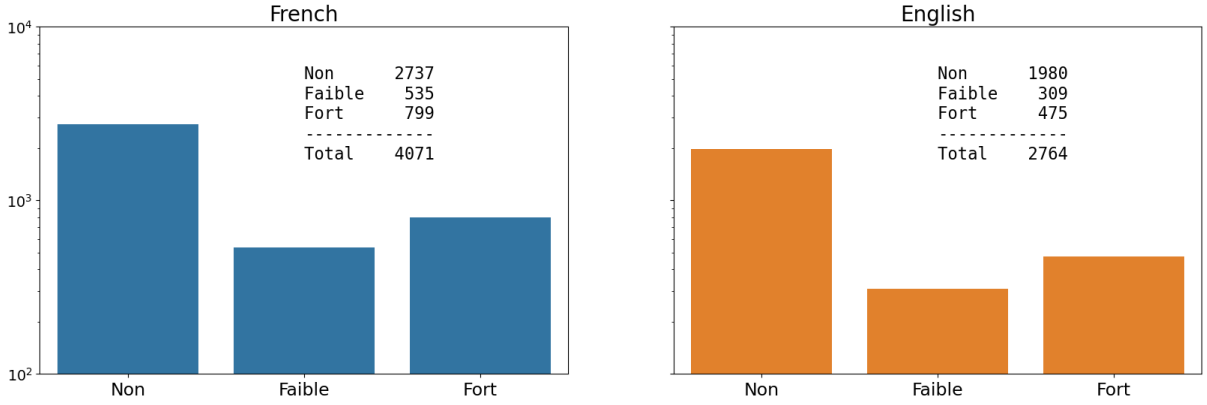


FIGURE 3.8
Class distribution within French and English for the host (on log scale).

3.6 PARTIAL RE-ANNOTATION AND INTER-ANNOTATOR AGREEMENT

Even if it has been defined in as many details as possible, the annotation regarding the structuring role of sentences obviously wears a subjective aspect. Two shows of the French dataset have therefore been re-annotated by someone else for comparison.

Table 3.5 shows the counts of the three classes in the two shows for the two annotators. If both are in order of magnitude equivalent, one can notice that the second one is more severe with the labels **Faible** and **Fort**.

Table 3.6 shows the kappa score between the two annotator for all the 848 sentences of the two shows and also when reducing on the 205 sentences of the host. Since it is defined that non-host sentences are always labelled **Non**, these a priori artificially pull upwards the kappa score, when they do not take part in the subjectivity of the task. One can however notice that kappa score does not reduce a lot when focusing on host sentences.

One aspect that the standard kappa score is probably missing is the continuity of the data. Indeed, through the annotation, one can often have the wish to annotate one and only one of two consecutive sentences, because the structure is clearly affected by the sentence pair, but none of them carry completely this role by itself. The solution in that case seems to chose to annotate one of the two sentences **Faible** or

	Main Annotator	Second Annotator
Non	778	794
Fort	39	31
Faible	31	23

TABLE 3.5
Comparison between the two annotators on the two shows that were annotated twice

	All Sentences (848)	Host Sentences (205)
Kappa score	0.7271	0.7047

TABLE 3.6
Kappa scores

Fort and the other one **Non**, but this obviously leads to a hesitation between the two. Therefore, in Figure 3.9, one can see the evolution of the kappa score if we count an agreement if the label appears at the same time $\pm N$ for $N = 0, 1, 2$ or 3. It can be seen that a tolerance of two sentences is sufficient to make the kappa score jump above 0.8, which is commonly considered as a good agreement. This might show the most important subjective part of the annotation and surely indicates that the structuring nature of the speech cannot be completely contained in individual sentences.

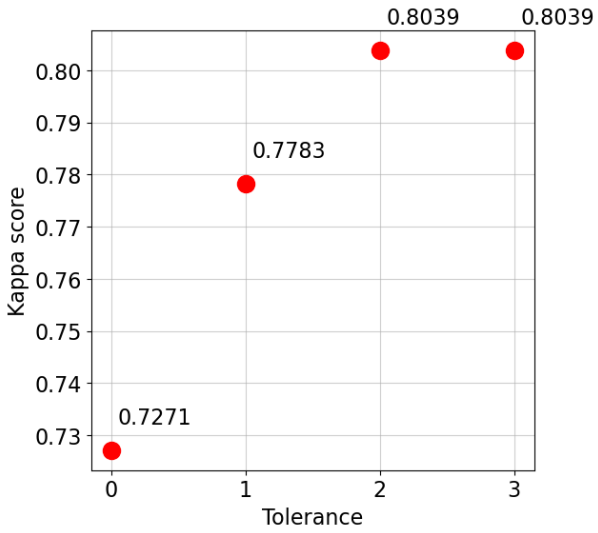


FIGURE 3.9
Evolution of the kappa score when counting an agreement
if there is only N sentences of mismatch between the two annotators.
For $N = 0, 1, 2$ or 3.

CHAPTER 4

EXPERIMENTS ON HOST DETECTION AND STRUCTURING QUESTIONS

> *You mentioned as objective two classifiers. How are they performing? Are they meeting your expectations?*

This chapter explores the tasks of detecting host utterances and structuring questions in podcast content by investigating the effectiveness of rule-based and BERT-based approaches in pinpointing host contributions. The aim is to enrich the broadcast transcripts to support their exploration.

All BERT fine-tuning experiments in this section are performed on an NVIDIA Tesla V100 GPU with 32 GB of memory. A colour-coding convention is used in the results tables to indicate the level of performance for each result. The key to this is provided in Table 4.1.

<0.5	**>0.6**	**>0.8**
Not satisfying	Encouraging	Good

TABLE 4.1
Colour code for precision, recall and F1-score performances.

4.1 HOST DETECTION: RULE-BASED METHOD VERSUS BERT-BASED MODELS

As the host plays a crucial role in the structure of the podcast, the detection of his or her utterances is a keystone in the appropriate enrichment of the audio content.

This section explores two different approaches to host detection in podcast episodes. The first is a rule-based method that uses the diarisation process to detect the host among the speakers in each episode. The second approach is to fine-tune BERT models with data that is annotated with host utterances in order to classify whether each sentence is spoken by the host or not. These model predictions may then be validated against the diarisation results to enhance accuracy.

4.1.1 RULE-BASED METHOD

Prior to this work, an exploratory study was conducted to detect the host of a given show, based on business rules and on its characteristics. Four criteria were identified and each speaker detected by diarisation in the show is given a score between 0 and 1 for each of these criteria. These criteria express the fact that the host is considered to be the speaker who ...

... **asks the most questions**: The speaker with the highest number of question marks in his/her transcript receive a score of 1. The other speakers get a score that correspond to the ratio between their number of question marks and that of the first speaker.

... **talks at the very beginning and very end of the show**: Here the first and last five sentences are taken into account. If only one speaker appears in both sets of sentences, he/she is attributed the score of 1. If several speakers appear in both sets, their score is their number of sentences in the two sets divided by 10.

... **talks on a regular basis during the show**: Here the start timestamps of the sentences are taken into account. The standard deviation of these timestamps is calculated for each speaker. The speaker with the highest deviation gets the score 1. The other speakers get a score that correspond to the ratio between their deviation and that of the first speaker.

... **often says the names of the guests**: Here the list of the names of the guests is provided. The occurrences of these names are counted in the transcript of each speaker. To do so, the library `spaCy`[1] is used to detect named entities. The entities detected as "`PER`" (for *person*) are compared to the list of the guests. An occurrence of a guest name is counted when the normalised Levenshtein distance between a detected mention and a name in the list of guest is greater than 0.6. Once again, the speaker with the most occurrences of guest names gets the score of 1, while the other a score proportional to this first speaker.

Finally, the four scores are summed and the speaker with the highest total score is considered as the host of the show. A visualisation of these four criteria for a typical example and their corresponding scores can be found in Figure 4.1.

With this method, the host is correctly identified in all podcasts of the dataset, up to diarisation errors. Indeed, even if the speaker ID is correctly selected, some sentences of this speaker are wrongly attributed to another one and vice versa. Therefore, sentences attributed to the host do not match entirely with the annotations. Table 4.2 shows the results for rule-based host detection, at sentence level, for the French and English test sets.

Language	Host precision	Host recall	Host f1-score	macro avg precision	macro avg recall	macro avg f1-score
FR	0,9506	0,9632	0,9569	0,9690	0,9731	0,97105
EN	0,9745	0,9636	0,9690	0,9796	0,9765	0,97805

TABLE 4.2
Host detection. Rule-based method on diarisation.
Sentence-wise results.

[1] `https://spacy.io/`

Speakers that ask the most questions:
SPEAKER_05 1.00
SPEAKER_01 0.52
SPEAKER_07 0.36
Speakers that are in the very beginning and end:
SPEAKER_05 1.00
Speakers with most extended replica curve:
SPEAKER_05 1.00
SPEAKER_03 0.89
SPEAKER_06 0.85
Speakers that name the most entities:
SPEAKER_05 1.00
SPEAKER_03 0.13
SPEAKER_06 0.13

The host is probably:
SPEAKER_05

Final ranking:
SPEAKER_05 4.00
SPEAKER_03 1.02
SPEAKER_06 0.98

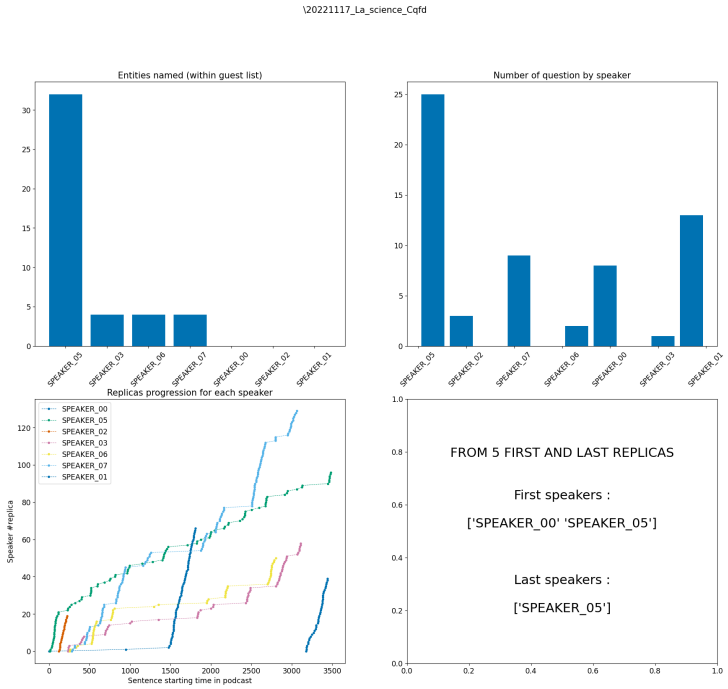


FIGURE 4.1
Visual representation of the four criteria for a typical show. In this case, it clearly (and truly) attributes the host role to the SPEAKER_05.

4.1.2 BERT MODELS

In these experiments, all BERT models of Table 2.1 are fine-tuned and evaluated (with precision, recall, and F1) separately on the French and English datasets, before combining them to assess the models' bilingual capabilities. Each sentence, labelled as *Animateur.ice*, is individually fed into the models to fine-tune this classification process, for two epochs.

MONOLINGUAL FINE-TUNING

Table 4.3 shows the results of host detection on the French test set for all models in terms of precision, recall, and F1-score. The precision of all models is overall commendable and the recall rates are satisfactory. The best precision (0,8131) and recall (0,8469) are achieved by `camembert-base`.

Performances on the English test set are presented in Table 4.4. The fine-tuned `bert-base-uncased` model achieves the highest precision, while `bert-base-cased` achieves the best recall. The performance of most models is robust, although `xlm-roberta-large` tends to overfit the **Non-host** class, and the recall of `camembert-base` is comparatively low.

An imbalance between precision and recall is noticeable in the results of both languages. This imbalance suggests that while models are mainly accurate in their positive predictions (precision), they vary in their ability to identify all relevant **Host** instances (recall), possibly overlooking some true host instances.

Model	Host precision	Host recall	Host f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,6847	0,5686	0,6213	0,7679	0,7363	0,7490
bert-large-cased	0,7102	0,5561	0,6238	0,7797	0,7365	0,7528
bert-base-uncased	0,7214	0,5810	0,6436	0,7890	0,7494	0,7649
bert-base-multilingual-cased	0,7735	0,5960	0,6732	0,8184	0,7660	0,7857
camembert-base	0,8131	0,6509	0,7230	0,8469	0,7980	0,8174
camembert-large	0,7920	0,6459	0,7115	0,8352	0,7919	0,8094
xlm-roberta-base	0,7633	0,6434	0,6982	0,8199	0,7851	0,7996
xlm-roberta-large	0,7785	0,5960	0,6751	0,8211	0,7669	0,7872

TABLE 4.3
Host detection with BERT. Finetuning on French dataset. Test on French.

Model	Host precision	Host recall	Host f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,7508	0,6751	0,7109	0,8100	0,7905	0,7991
bert-large-cased	0,7276	0,6583	0,6912	0,7949	0,7774	0,7851
bert-base-uncased	0,7766	0,6331	0,6975	0,8169	0,7783	0,7933
bert-base-multilingual-cased	0,7576	0,6303	0,6881	0,8063	0,7728	0,7861
camembert-base	0,7715	0,5770	0,6603	0,8055	0,7527	0,7710
camembert-large	0,7377	0,6303	0,6798	0,7958	0,7681	0,7794
xlm-roberta-base	0,7687	0,6611	0,7108	0,8172	0,7888	0,8006
xlm-roberta-large	0	0	0	0,3522	0,5000	0,4133

TABLE 4.4
Host detection with BERT. Finetuning on English dataset. Test on English.

BILINGUAL FINE-TUNING

Next, the same models are fine-tuned on the training sets of both languages and tested on each language separately. Tables 4.5 and 4.6 show the results for French and English test sets, respectively.

When testing on French (Table 4.5), the best precision is obtained by `camembert-large` and the best recall by `camembert-base`. Precision remains consistent with the previous fine-tuning on the French training set only, but recall improves overall, with the exception of `xlm-roberta-large`, which drops to zero. When testing on English (Table 4.6), the best precision is reached by `camembert-base` and best recall by `bert-large-cased`. A slight reduction in precision is observed alongside a minor improvement in recall, except for `xlm-roberta-large`, where performance is again suboptimal.

CONCLUSIONS

The conclusions drawn from this analysis highlight that the predictions for host detection, derived from comparing BERT prediction for each sentence with the ground truth, do not initially factor in the diarisation results. However, incorporating diarisation by identifying the speaker most frequently associated with host-labeled by BERT sentences can refine host detection accuracy. This approach consistently aligns the model's predictions with the correct host across various tests, except in instances involving `xlm-roberta-large`, where the **Host** class is not identified. This method demonstrates that while sentence-level performance heavily relies on diarisation, it ensures consistency in host identification, as evidenced by the results matching those in Table 4.2.

Model	Host precision	Host recall	Host f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,7036	0,5860	0,6395	0,7803	0,7478	0,7610
bert-large-cased	0,6877	0,6259	0,6554	0,7775	0,7609	0,7683
bert-base-uncased	0,7085	0,6060	0,6532	0,7857	0,7573	0,7692
bert-base-multilingual-cased	0,7946	0,5885	0,6762	0,8284	0,7664	0,7888
camembert-base	0,7803	0,6733	0,7229	0,8332	0,8019	0,8153
camembert-large	0,8056	0,6509	0,7200	0,8430	0,7966	0,8152
xlm-roberta-base	0,7630	0,6584	0,7068	0,8219	0,7917	0,8046
xlm-roberta-large	0	0	0	0,3659	0,5000	0,4226

TABLE 4.5
Host detection with BERT. Finetuning on French and English datasets. Test on French.

Model	Host precision	Host recall	Host f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,7185	0,7507	0,7342	0,806	0,8137	0,8096
bert-large-cased	0,6951	0,7535	0,7231	0,7940	0,8074	0,8000
bert-base-uncased	0,7252	0,7171	0,7211	0,8035	0,8016	0,8025
bert-base-multilingual-cased	0,7327	0,6835	0,7072	0,8018	0,7894	0,7951
camembert-base	0,7444	0,6527	0,6955	0,8029	0,7793	0,7893
camembert-large	0,7028	0,7087	0,7057	0,7901	0,7915	0,7908
xlm-roberta-base	0,6992	0,7227	0,7107	0,7906	0,7961	0,7932
xlm-roberta-large	0	0	0	0,3522	0,5000	0,4133

TABLE 4.6
Host detection with BERT. Finetuning on French and English datasets. Test on English.

4.2 STRUCTURING QUESTION DETECTION

4.2.1 BERT MODELS: THREE LABEL CLASSIFIER

This section focuses on conducting supervised fine-tuning of BERT models to determine whether a sentence contributes structurally to a discourse. The models are trained for five epochs, first using separate French and English training sets (monolingual fine-tuning), with the F1-score serving as the performance metric. Each sentence, labelled as *Structurante*, is individually fed into the models to facilitate this classification process. Structuring sentence classification involves determining the level of structuring a sentence provides to a conversation, labelled as **Fort** (Strong), **Faible** (Weak), or **Non** (None), with a primary focus on the **Fort** category as it is of major practical interest.

MONOLINGUAL FINE-TUNING

Tables 4.7 and 4.8 present the results obtained on the French and English testing sets, respectively. Precision, recall and F1-score are indicated for macro average and for the label **Fort**.

Looking at the results for the French dataset in Table 4.7, `bert-base-multilingual-cased` achieves the highest precision for **Fort** class, and `xlm-roberta-base` the highest recall. However, apart from the overfitting models, none of the results is significantly higher than another, except to some extent for the recall of `xlm-roberta-base`, and none of them can be considered satisfactory. It can also be noticed that the large base models perform better than the tested large equivalents. As for the macro averages, results are higher, but are pulled upwards by the over-represented (and easy) **Non**.

Model	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,4667	0,4730	0,4698	0,6755	0,5805	0,6093
bert-large-cased	0,4138	0,4865	0,4472	0,7025	0,5575	0,5784
bert-base-uncased	0,4432	0,5270	0,4815	0,6577	0,5726	0,5868
bert-base-multilingual-cased	0,4868	0,5000	0,4933	0,6735	0,6037	0,6278
camembert-base	0,4737	0,4865	0,4800	0,6694	0,6050	0,6281
camembert-large	0	0	0	0,3014	0,3333	0,3166
xlm-roberta-base	0,4245	0,6081	0,5000	0,7290	0,6230	0,6308
xlm-roberta-large	0	0	0	0,4688	0,3459	0,3427

TABLE 4.7
Structuring questions detection. Three label classifier.
Fine-tuning on French dataset. Test on French.

Model	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,5070	0,5217	0,5143	0,6509	0,6010	0,6195
bert-large-cased	0,5616	0,5942	0,5775	0,5923	0,5934	0,5924
bert-base-uncased	0,4730	0,5072	0,4895	0,5685	0,5831	0,5755
bert-base-multilingual-cased	0,4925	0,4783	0,4853	0,6184	0,5467	0,5689
camembert-base	0,5185	0,6087	0,5600	0,6943	0,5593	0,5631
camembert-large	0,5469	0,5072	0,5263	0,6416	0,5475	0,5736
xlm-roberta-base	0,4742	0,6667	0,5542	0,4809	0,5465	0,5083
xlm-roberta-large	0	0	0	0,3052	0,3333	0,3186

TABLE 4.8
Structuring questions detection. Three label classifier.
Fine-tuning on English dataset. Test on English.

For the English dataset, in Table 4.8, `bert-large-cased` obtains the best precision for **Fort** class, and `xlm-roberta-base` the best recall. In that case, the large models perform better than their base equivalent (except for the overfitting `xlm-roberta-large`). The quality of the results is a bit more variable between the models, but none gets a satisfying precision and only `camembert-base` and `xlm-roberta-base` get an acceptable recall. Again, the macro averages are higher, but pulled upwards by the **Non** class.

In appendix, results for cross-lingual evaluation are shown, i.e. fine-tuning on French data with test on English and vice versa. These can be seen in Tables 8.1 and 8.2. This gives a first insight on how the models can generalise to other languages. One can see that, in general, English performs better on French fine-tuning than the opposite, but most of the models present difficulties to generalise to the other language.

CONCLUSIONS The results show that the **Fort** class performs better in the English dataset. While this superior performance is in line with expectations for most models, given the dominant role of English pretrained model development, it is notably surprising for the French-focused model `camembert`. A pattern of overfitting to the **Non** class is evident among the larger models, preventing them from effectively predicting the **Faible** or **Fort** categories, as seen with `xlm-roberta-large` in both languages and `camembert-large` in the French dataset, but not in English. Typically, `bert-base` models with casing (cased) show better performance than their uncased counterparts, probably due to the retention of case-sensitive information in the text.

Model	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,5385	0,4730	0,5036	0,6767	0,5860	0,6198
bert-large-cased	0,3976	0,4459	0,4204	0,6165	0,5236	0,5369
bert-base-uncased	0,4521	0,4459	0,4490	0,6578	0,5758	0,6035
bert-base-multilingual-cased	0,5373	0,4865	0,5106	0,6770	0,5908	0,6226
camembert-base	0,5059	0,5811	0,5409	0,7188	0,6175	0,6400
camembert-large	0	0	0	0,3014	0,3333	0,3166
xlm-roberta-base	0,4211	0,5405	0,4734	0,7456	0,5931	0,6138
xlm-roberta-large	0	0	0	0,3014	0,3333	0,3166

TABLE 4.9
Structuring questions detection. Three label classifier.
Fine-tuning on French and English datasets. Test on French.

Model	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,5161	0,4638	0,4885	0,5895	0,5894	0,5882
bert-large-cased	0,4337	0,5217	0,4737	0,5510	0,5383	0,5359
bert-base-uncased	0,5000	0,5072	0,5036	0,5874	0,5843	0,5858
bert-base-multilingual-cased	0,6056	0,6232	0,6143	0,6361	0,6229	0,6285
camembert-base	0,5667	0,4928	0,5271	0,6554	0,5632	0,5953
camembert-large	0	0	0	0,3052	0,3333	0,3186
xlm-roberta-base	0,5658	0,6232	0,5931	0,6600	0,5763	0,5869
xlm-roberta-large	0	0	0	0,3052	0,3333	0,3186

TABLE 4.10
Structuring questions detection. Three label classifier.
Fine-tuning on French and English datasets. Test on English.

BILINGUAL FINE-TUNING

The next idea is to see if training all models on both datasets (bilingual fine-tuning) can improve the performance on each language individually, the intuition being that more data would enhance the fine-tuning process and that the identified features raised by the classifiers are already at a minimum level independent of the language. The models are then trained under the same conditions as before, except that they are now given both French and English training and validation sets. The results are presented in Tables 4.9, respectively 4.10, for French and English performances.

As before, the results are globally better for English than for French for almost all models (except `bert-base-cased`) and the large models of `xlm-roberta` and `camembert` (this time for both languages) have no **Fort** prediction.

For French, in Table 4.9, `bert-base-cased` obtains the highest precision and `camembert-base` the highest recall. Again, the base models perform better than their large equivalent.

For English, in Table 4.10, `bert-base-multilingual-cased` obtains the highest precision and recall, tied with `xlm-roberta-base` for recall.

CONCLUSIONS For the bilingual fine-tuning, the gaps between the different results begin to widen as the performances of the base models increase that of the large models decrease. It can be noticed that the results for `bert-base-multilingual` for English are the first to be homogeneous and satisfactory in any terms (precision, recall, F1-score for class **Fort** or macro average).

4.2.2 BERT MODELS: BINARY CLASSIFICATION

As the initial experiments did not yield particularly strong results, the strategy shifts towards simplifying the task to a binary classification, focusing on distinguishing between the **Non** and **Fort** classes.

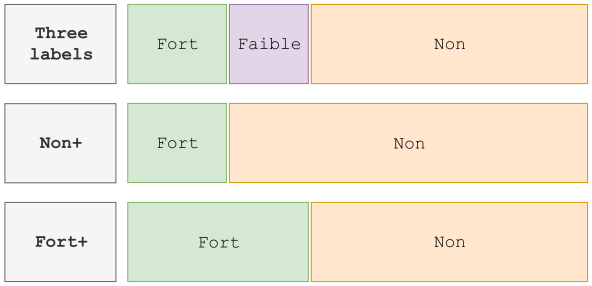


FIGURE 4.2
Two ways of relabelling the Faible data (not to scale).

In order to adapt to this binary framework, sentences previously annotated as **Faible** are relabelleded in two ways. First, they are reclassified as **Non** to emphasize the distinctiveness of the **Fort** class. This relabelling is referred to as the "Non+" configuration. Second, sentences annotated as **Faible** are relabelled as **Fort** to achieve a more balanced distribution between the two classes. This second relabelling is referred to as the "Fort+" configuration.A schematic view of this relabelling is shown in Figure 4.2. The original multiclass annotation configuration is referred to as "three label classifier".

CONFIGURATION *Non+*

Model	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,0345	0,0270	0,0303	0,4922	0,4938	0,4928
bert-large-cased	0,4769	0,4189	0,4460	0,7234	0,6975	0,7095
bert-base-uncased	0,0411	0,0405	0,0408	0,4956	0,4956	0,4956
bert-base-multilingual-cased	0,0179	0,0135	0,0154	0,4836	0,4874	0,4853
camembert-base	0,0375	0,0405	0,0390	0,4937	0,4932	0,4934
camembert-large	0	0	0	0,4753	0,5000	0,4873
xlm-roberta-base	0,0312	0,0270	0,0290	0,4905	0,4917	0,4910
xlm-roberta-large	0	0	0	0,4753	0,5000	0,4873

TABLE 4.11
Structuring questions detection. "Non+" configuration.
Fine-tuning on French dataset. Test on French.

Tables 4.11, 4.12, 4.13 and 4.14 present the results of the four same experiments, i.e. French and English performances with fine-tuning on the same language or on both of them, for the "Non+" configuration.

For French, in Table 4.11, `bert-large-cased` achieves the highest precision and recall for the **Fort** class. In fact, the performance of all other model is practically at zero.

For English, in Table 4.12, the best precision is `xlm-roberta-large`, but with a very low recall. The second best precision is `bert-base-cased`, which also achieves a recall within the average of that of other models. The best recall is however obtained by `bert-large-cased`.

Model	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,6667	0,4928	0,5667	0,8182	0,7389	0,7720
bert-large-cased	0,5692	0,5362	0,5522	0,7706	0,7558	0,763
bert-base-uncased	0,6207	0,5217	0,5669	0,796	0,7512	0,7715
bert-base-multilingual-cased	0,5660	0,4348	0,4918	0,7661	0,7073	0,7324
camembert-base	0,5200	0,3768	0,4370	0,7414	0,6779	0,7039
camembert-large	0	0	0	0,4714	0,5000	0,4853
xlm-roberta-base	0,4643	0,3768	0,4160	0,7135	0,6752	0,6921
xlm-roberta-large	0,8750	0,1014	0,1818	0,9117	0,5503	0,5774

TABLE 4.12
Structuring questions detection. “Non+” configuration.
Fine-tuning on English dataset. Test on English.

Model	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,4848	0,4324	0,4571	0,7277	0,7043	0,7152
bert-large-cased	0	0	0	0,4753	0,5000	0,4873
bert-base-uncased	0,5000	0,4324	0,4638	0,7353	0,7050	0,7189
bert-base-multilingual-cased	0,4921	0,4189	0,4526	0,7310	0,6982	0,7131
camembert-base	0,5211	0,5000	0,5103	0,7476	0,7380	0,7427
camembert-large	0,4754	0,3919	0,4296	0,7220	0,6847	0,7013
xlm-roberta-base	0,4605	0,4730	0,4667	0,7165	0,7221	0,7192
xlm-roberta-large	0	0	0	0,4753	0,5000	0,4873

TABLE 4.13
Structuring questions detection. “Non+” configuration.
Fine-tuning on French and English datasets. Test on French.

When fine-tuning on both languages and testing on French, in Table 4.13, `camembert-base` gets the best precision and recall for the **Fort** class.

When testing on English, in Table 4.14, it also gets the highest precision, but `bert-base-cased` gets the highest recall. On a multilingual perspective, `bert-base-multilingual` yields better results than `xlm-roberta-base` in terms of precision on the **Fort** class, and inversely for recall, for both languages. However, these results can be considered as close to each other.

In Tables 8.3, respectively 8.4, in appendix, the cross-lingual evaluations can be seen.

CONFIGURATION *Fort+*

Tables 4.15, 4.16, 4.17 and 4.18 present the results of the four same experiments, i.e. French and English performances with fine-tuning on the same language or on both of them, for the “Fort+” configuration.

Specifically, when fine-tuning is conducted on the corresponding language of the test set, French models (Table 4.15) have significant difficulties in accurately predicting the **Fort** class, with performance metrics nearing zero for most models except for `bert-large-cased` variants, which show improved outcomes compared to the “Non+” configuration.

In the case of English (Table 4.16), `bert-large-cased` obtains the best precision and `bert-base-uncased` the best recall. It can be noticed that “Fort+” results for most of the cases are substantially higher than

Model	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,5294	0,5217	0,5255	0,7502	0,7468	0,7485
bert-large-cased	0	0	0	0,4714	0,5000	0,4853
bert-base-uncased	0,5254	0,4493	0,4844	0,7462	0,7123	0,7278
bert-base-multilingual-cased	0,5882	0,4348	0,5000	0,7773	0,7082	0,7369
camembert-base	0,6000	0,4783	0,5323	0,7844	0,7295	0,7535
camembert-large	0,5217	0,3478	0,4174	0,7415	0,6643	0,6941
xlm-roberta-base	0,5769	0,4348	0,4959	0,7716	0,7077	0,7346
xlm-roberta-large	0	0	0	0,4714	0,5000	0,4853

TABLE 4.14
Structuring questions detection. “Non+” configuration.
Fine-tuning on French and English datasets. Test on English.

Model	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,0609	0,0490	0,0543	0,4812	0,4845	0,4825
bert-large-cased	0,7073	0,6084	0,6541	0,8333	0,7909	0,8102
bert-base-uncased	0,0583	0,0490	0,0532	0,4797	0,4827	0,4810
bert-base-multilingual-cased	0,0648	0,0490	0,0558	0,4834	0,4871	0,4846
camembert-base	0,0603	0,0490	0,0541	0,4809	0,4842	0,4822
camembert-large	0,0583	0,0490	0,0532	0,4797	0,4827	0,4810
xlm-roberta-base	0,0472	0,0420	0,0444	0,4735	0,4762	0,4748
xlm-roberta-large	0	0	0	0,4522	0,5000	0,4749

TABLE 4.15
Structuring questions detection. “Fort+” configuration.
Fine-tuning on French dataset. Test on French.

in the “Non+” configuration, even if the poor French fine-tuning results make the comparison not really pertinent.

For the fine-tuning on both languages, the results are also better than in the “Non+” configuration. For French (Table 4.17), `bert-base-cased` has the highest precision and `xlm-roberta-base` the highest recall.

For English (Table 4.18), `bert-base-multilingual` has the highest precision and `bert-large-cased` the highest recall. For both languages, `bert-base-multilingual` also gets the highest multilingual precision when `xlm-roberta-base` the highest recall.

In Tables 8.5, respectively 8.6, in appendix, the cross-lingual evaluations can be seen.

4.3 ANALYSIS AND DISCUSSION

4.3.1 HOST DETECTION

Regarding the host detection, the first important observation is the very good results of the rule-based approach (cf. Table 4.2). Indeed, with a standard tool of diarisation and a few simple rules, it is possible to accurately detect the sentences uttered by the host, both in terms of precision and recall. However, such a good result should be taken with a pinch of salt, since the method suffers from many weaknesses. First of all, the four chosen criteria seem arbitrary. They are, indeed, all chosen on the basis of some intuition

Model	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,7447	0,6863	0,7143	0,8580	0,8323	0,8445
bert-large-cased	0,7717	0,6961	0,7320	0,8720	0,8385	0,8543
bert-base-uncased	0,7282	0,7353	0,7317	0,8519	0,8550	0,8534
bert-base-multilingual-cased	0,7326	0,6176	0,6702	0,8489	0,7984	0,8212
camembert-base	0,7500	0,5588	0,6404	0,8551	0,7708	0,8059
camembert-large	0,6875	0,6471	0,6667	0,8276	0,8100	0,8185
xlm-roberta-base	0,7215	0,5588	0,6298	0,8408	0,7695	0,7999
xlm-roberta-large	0,5631	0,5686	0,5659	0,7616	0,7640	0,7628

TABLE 4.16
Structuring questions detection. “Fort+” configuration.
Fine-tuning on English dataset. Test on English.

Model	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,7455	0,5734	0,6482	0,8507	0,7764	0,8079
bert-large-cased	0,7273	0,5594	0,6324	0,8409	0,7686	0,7992
bert-base-uncased	0,7288	0,6014	0,659	0,8437	0,7889	0,8132
bert-base-multilingual-cased	0,7304	0,5874	0,6512	0,8438	0,7822	0,8091
camembert-base	0,7227	0,6014	0,6565	0,8406	0,7885	0,8117
camembert-large	0	0	0	0,4522	0,5000	0,4749
xlm-roberta-base	0,7164	0,6713	0,6931	0,8409	0,8216	0,8309
xlm-roberta-large	0	0	0	0,4522	0,5000	0,4749

TABLE 4.17
Structuring questions detection. “Fort+” configuration.
Fine-tuning on French and English datasets. Test on French.

related to the specificities of the particular dataset of this thesis and probably cannot be generalised to all kinds of podcast. Furthermore, some criteria are likely to be correlated between them. For example, the standard deviation of speakers utterances and the list of those who speak at the very beginning and at the very end. It is clear that someone who fulfils the latter criteria will also have a high standard deviation. Therefore, the two do not really provide independent information. In addition to that, one of the criteria requires the availability of a guest list. This is a strong constraint that binds the process to human guidance and compromises its independence, not to mention that, in practice, the detection of the host might itself be an appropriate method to facilitate the identification of guest names. This should therefore not intervene in the host detection. Finally, it should be emphasised that the rule-based approach is entirely dependent on the performance of the diarisation; however, given the good performance already offered by standard tools, this is not (any longer) a major issue. That being said, the four criteria can however be kept that way in an industrial context for an application to podcasts that are not too dissimilar to those of the present work.

For their part, BERT models yield interesting results. In particular, some have already quite high precision and recall, for example `camembert-base` in the French fine-tuning in Table 4.3, but it is true on a very broad point of view in Tables 4.3, 4.4, 4.5 and 4.6. These are encouraging results given that the models only take in each sentence independently and lack a lot of context such as the other sentences around it, diarisation information, timestamps, etc. It seems that there is some intrinsic information that can be identified from the sentences of the host themselves, and also that these are not language dependant, since fine-tuning on French and English simultaneously improves the performance for both languages, when

Model	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,6449	0,6765	0,6603	0,8074	0,8211	0,8141
bert-large-cased	0,7103	0,7451	0,7273	0,8433	0,8585	0,8507
bert-base-uncased	0,6863	0,6863	0,6863	0,8287	0,8287	0,8287
bert-base-multilingual-cased	0,7195	0,5784	0,6413	0,8407	0,7788	0,8059
camembert-base	0,7011	0,5980	0,6455	0,8323	0,7873	0,8077
camembert-large	0	0	0	0,4578	0,5000	0,4780
xlm-roberta-base	0,701	0,6667	0,6834	0,8352	0,8202	0,8275
xlm-roberta-large	0	0	0	0,4578	0,5000	0,4780

TABLE 4.18
Structuring questions detection. “Fort+” configuration.
Fine-tuning on French and English datasets. Test on English.

comparing Table 4.5 with 4.3 for French and Table 4.6 with 4.4 for English.

However, comparing all these results with Table 4.2, it seems unwise to estimate that the proper nature of host sentences could be more discernable for a BERT model than vocal features for a diarisation tool. The latter should be seen as a necessary step that will certainly become more accurate over time, and the discrimination of the host among speakers the best way to find his utterances. BERT models are a reliable factor, whatever the language (French or English) or BERT model (except for some the `xlm-roberta-large` overfitting). For the selection of the host among the speakers, these classifiers should certainly be used, as unique criteria as proposed at the end of Section 4.1.2 or in a possible mix with some criteria of the rule-based approach.

4.3.2 STRUCTURING QUESTION DETECTION

Section 3.6 showed that the manual annotation of the structuring nature of sentences is rather subjective. This explains probably the poor results that are obtained, in particular for the three label classifier in Tables 4.7, 4.8, 4.9 and 4.10. In addition, sentences annotated as structuring are often a series of consecutive sentences rather than a single structuring sentence containing all the information. Since BERT classifiers are fine-tuned to predict the structuring nature of each sentence individually, it is not surprising that it fails having good results in general. Not only the classifier has to make a choice that is partially subjective, it also does it without context. In the annotation process, since the structuring nature of a sentence has been defined as a sentence that induces a change of topic, the fact that a different topic is covered after a sentence uttered by the host probably induced the annotator for a positive annotation, whereas the classifier cannot have this kind of information.

From the point of view of the different models, none of them stands out from the crowd by systematically performing better than the others, and all of them are often in the same order of magnitude. This can be observed in all cases, i.e. between Tables 4.7, 4.8, 4.9 and 4.10 for the three label classifier, between Tables 4.11, 4.12, 4.13 and 4.14 for the “Non+” configuration and between Tables 4.15, 4.16, 4.17 and 4.18 for the “Fort+” configuration. However, an exception should be made for the large models. Indeed, even if they sometimes achieve the best results for some situations, they all at least once overfit to the **Non** class, resulting in no **Fort** prediction. This unstable behaviour makes them not conducive in practice, not to mention that their performances are not even much better than their base equivalent.

Regarding the binary classifiers, it clearly appears that the “Fort+” configuration leads to better results than the “Non+” one, comparing respectively Tables 4.11 with 4.15, 4.12 with 4.16, 4.13 with 4.17 and 4.14 with 4.18. For a similar situation, the difference is in general significant. This suggests that the

difference between the **Faible** and **Fort** classes is probably more subtle than between each of them with the **Non** class and that the three-classes classifier is confusing **Faible** and **Fort** more often. However, this might also signify that, in the "Fort+" configuration, the **Fort** class is weakened and that it only yields better results by being more tolerant, since the **Faible** sentences are accepted as **Fort**. In other words, the "Fort+" configuration classification is an easier task, which makes the results better, but these are less interesting in practice.

In the two binary configurations, the very poor performances of the French fine-tunings in Tables 4.11 and 4.15, must be mentioned. Except for bert-large-cased, their results are disproportionally lower than any other result in Section 4.2.2. This behaviour is unexplained. One cannot blame an overfitting problem, since all models predict a proportion of **Fort** sentences in accordance with the proportion in the ground truth. It is even more surprising that this is not observed at all in the three label classifier (Table 4.7) and that in the binary configuration, when the English fine-tuning data is added (Tables 4.13 and 4.17), the performances on the same French test set climb back up to the norm.

To better compare these different performances between BERT models and configurations, the results are summarised in Figure 4.3a for precision and 4.3b for recall. They also represent the cross-lingual evaluation in the three configurations for more global comparison.

One can notice the overall better performance on the English test set compared to the French one. For the three label classifier, this is particularly noticeable when comparing the monolingual fine-tuning in Tables 4.7 and 4.8, where the English case yields better results despite a smaller amount of fine-tuning data. The unexplained poor performance of the French fine-tuning in binary configuration, in Tables 4.11 and 4.15, make the comparison with English cases inappropriate.

More interestingly, when the two languages are combined for fine-tuning, the performance increase in both languages (except for some large models, which suggests even more against using them). It is true for French comparing Tables 4.7 and 4.9 for the three label classifier, and Tables 4.11 and 4.13, respectively Tables 4.15 and 4.17, for the "Non+", respectively "Fort+", configurations. The same goes for English, comparing Tables 4.8 and 4.10 for the three label classifier, and Tables 4.12 and 4.14 for the "Non+", with the exception of the "Fort+", configurations in Tables 4.16 and 4.18. This indicates that the models seem to be able to be extended to other languages and that the information that matters for capturing the structuring nature of a sentence is language-independent. In the perspective of the multilingual extension of the experiment, comparing the two multilingual-minded models tested shows that `bert-base-multilingual` always gives better results than `xlm-roberta-base`, from the point of view of the **Fort** precision, while the opposite is in generally true for **Fort** recall. In practice, for the application to the EBU prototype, the precision should be preferred, as it is preferable that users miss a structuring question rather than to be shown many false positives. Therefore, `bert-base-multilingual` will be used in the experiments of Section 5.

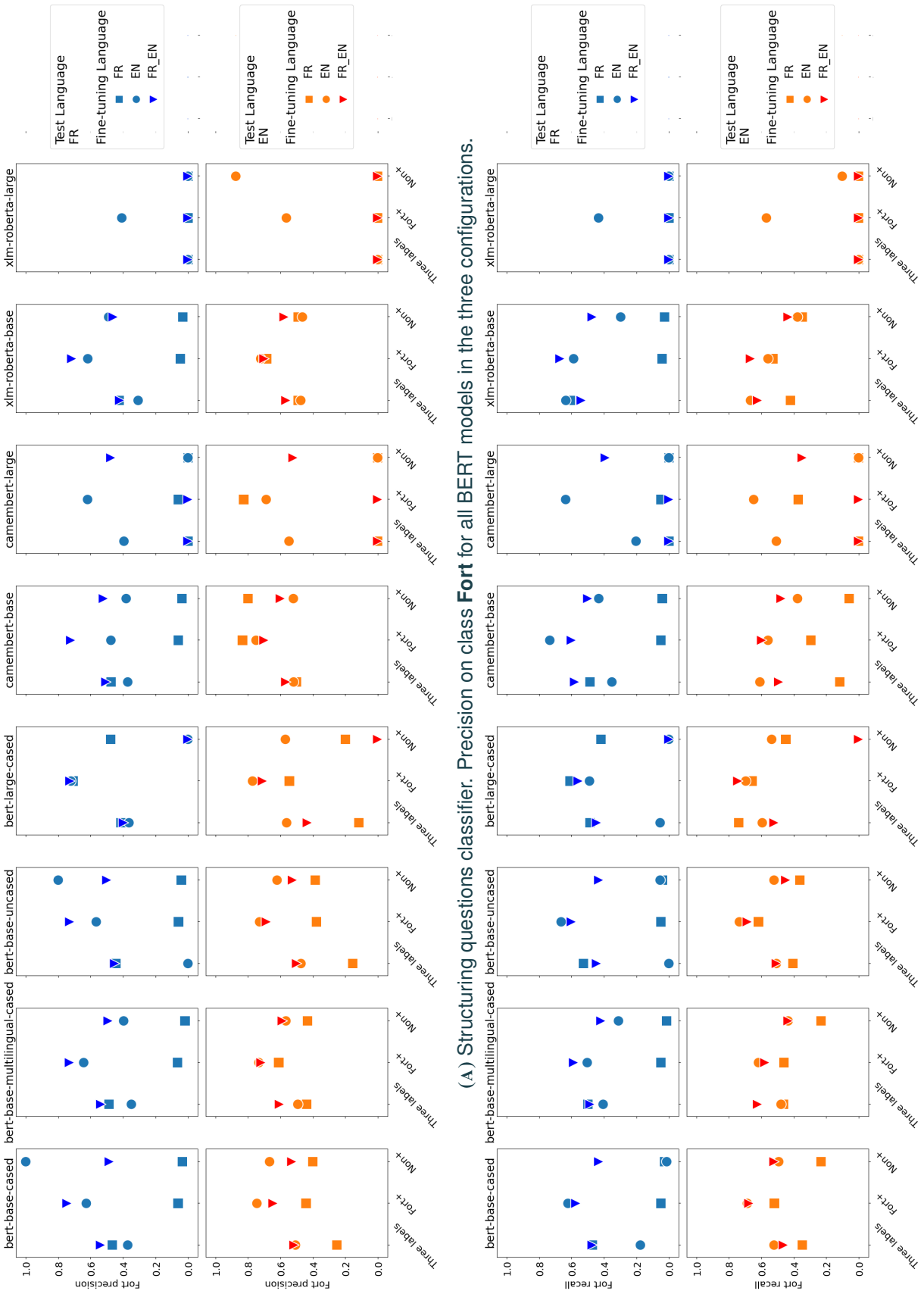
4.4 CONCLUSIONS

Overall, no particular BERT model could be identified that is better than another for both host and structuring question detection. However, the frequent risk of overfitting of large models has been identified, which seems to recommend against their use in practice.

With respect to host detection, French and English models (monolingual fine-tuning) seem to perform similarly in terms of precision. However, the recall is generally better for English than for French. Nevertheless, both languages give sufficiently satisfactory results to provide a credible alternative to the rule-based method, especially if BERT predictions are used as a rule in diarisation. Only good precision is required.

From the perspective of structuring question detection, English performances are globally better than French, particularly if fine-tuning is performed on one language only. However, for both, the results improve with bilingual fine-tuning. Binary classifiers have been found to be easier tasks, particularly the "Fort+" configuration. However, in these situations, fine-tuning on French only has shown an unexplained counter-performing behaviour for almost all models.

The method proposed so far has evident areas for improvement. Firstly, fine-tunings and predictions are made without any context information, which could be good guidance and is already easily available. This will be a matter of focus in Section 6. Secondly, the wide range of languages that the industrial context of this thesis requires to be handled has not been taken into account until now. This will be explored in Section 5.



(A) Structuring questions classifier. Precision on class **Fort** for all BERT models in the three configurations.

(B) Structuring questions classifier. Recall on class **Fort** for all BERT models in the three configurations.

FIGURE 4.3

CHAPTER 5

EXPERIMENTS ON LANGUAGE VARIABILITY

> *So yes, let's talk about the EBU context of this thesis. Indeed, to make your tool available for all Members, it has to be multilingual. How did you address this challenge?*

As part of the insertion in the context of the EBU radio tool *Meta Radio* for its Members, it is necessary to have, at least an insight, of how it performs on other languages in its current state and if it is possible to perform a multilingual fine-tuning. The need is at this point to have some data in other languages, with the problematic that it is impossible to ask for the assistance of experts of these new languages. This could not have been a problem for obtaining some transcriptions of podcasts, since the automatic speech-to-text tools are supposed to be reliable enough, but rather for being able to annotated them, in particular the structuring class.

5.1 AUTOMATIC TRANSLATION AND EVALUATION

Thus, the first step of the strategy involves automatically translating the French and English podcasts not only between these two languages but also into eighteen additional languages, aiming to compile a dataset that encompasses twenty languages, reflective of the diversity within the EBU membership. The chosen languages span several linguistic families: Indo-European languages like Greek, Albanian, and also subfamilies such as Slavic languages with Slovenian, Bulgarian, Polish, and Ukrainian; the Baltic languages with Latvian; Western Germanic languages including English, German, and Dutch; Scandinavian languages with Swedish and Norwegian; Italic languages such as French, Romanian, Italian, and Spanish (Castilian). The selection is rounded off with the Uralic languages, Hungarian and Finnish, along with Turkish and Arabic, to ensure a broad and representative linguistic spectrum.

Figure 5.1 shows the different languages, as well as the provider used for the automatic translation. DeepL[1] is used for almost all languages, except for Albanian and Arabic, that are not supported by DeepL and are therefore translate with Google Translate[2]. The requests are sent through an internal tool of the EBU called Eurovox[3].

[1] `https://www.deepl.com/translator`
[2] `https://translate.google.com/`
[3] `https://tech.ebu.ch/eurovox`

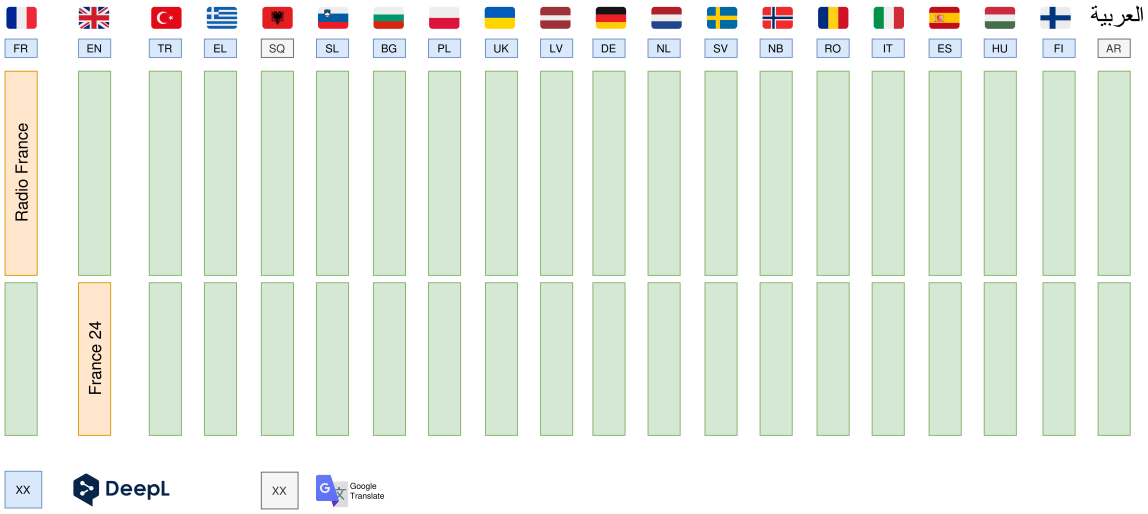


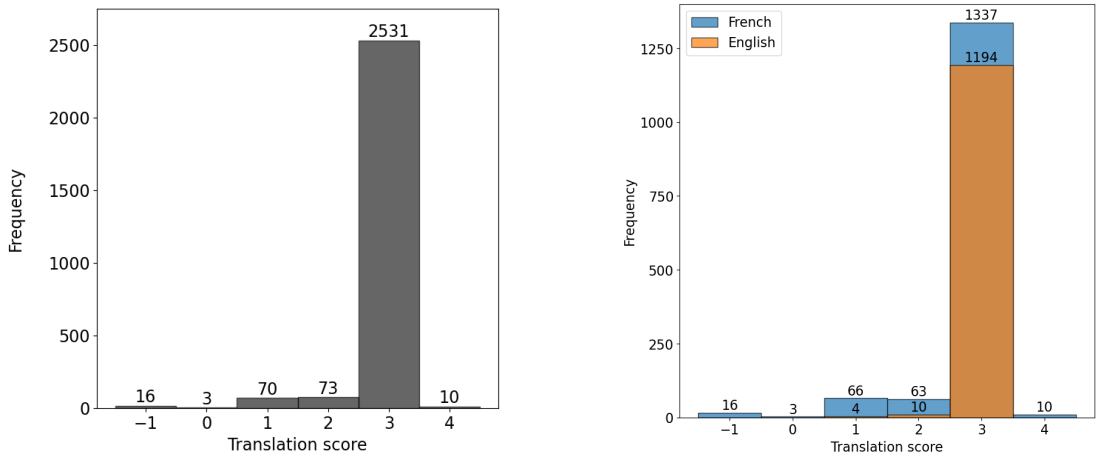
FIGURE 5.1
Automatic augmentation of the dataset in 20 languages.

5.1.1 EVALUATION

This new augmented dataset, with twenty languages, needs to be, at least partially, evaluated in terms of translation quality. Due to the substantial size of the initial dataset and the challenge of securing experts for the diverse range of selected languages, the evaluation is conducted solely on the German translations of the French and English test sets. Each translated sentence is compared to its original and given a score corresponding to four criteria.

- **Score 4:** *Enhanced Translation* - The translation improves upon the original, possibly by correcting typos, time concordance, punctuation, or resolving homonymic confusion.
- **Score 3:** *Accurate Translation* - The translation is reliable and faithful to the original sentence.
- **Score 2:** *Acceptable with Minor Issues* - The translation has slight inaccuracies that do not alter the sentence's overall meaning or structure, such as unnecessary changes to abbreviations or proper nouns.
- **Score 1:** *Problematic Translation* - The translation has significant issues affecting the sentence's meaning or structure, like incorrect pronoun or determiner usage, incomplete rendering, or missing nuanced expressions.
- **Score 0:** *Inaccurate Translation* - The translation is completely incorrect, nonsensical, or the sentence remains untranslated.
- **Score -1:** *Irrelevant for Evaluation* - The original sentence has major issues, rendering the translation's evaluation irrelevant.

This scoring framework aims to systematically gauge the quality of translations within the dataset.The results of the evaluation can be seen in Figure 5.2a, respectively 5.2b, for the translation of the two test sets, respectively for each initial language separately. It clearly appears that the translation is for an overwhelming majority excellent. There is only the order of 5 % of the sentences that suffer from problems, the half of which seeming negligible. Interestingly, the errors almost all come from the translation from French.



(A) Combined French and English test sets.

(B) Individual French and English test sets.

FIGURE 5.2
Evalutation of German automatic translation.

5.2 STRUCTURING TESTING ON GERMAN TRANSLATION

After the evaluation of its quality, the new German test set is used to test the performances of `bert-base-multilingual-cased` fine-tuned in several situations. It is fine-tuned over two epochs respectively on the initial French and English datasets separately (i.e. without translated sentences), then on the merge of the two, plus on the German dataset and finally on the merge of the three languages. In addition to that, the model `bert-base-german-cased` is also fine-tuned on the German dataset to evaluate its performance on the German test set.

Tables 5.1, respectively 5.2 and 5.3, show the results of the fine-tuning and testing on the German test set for the three label classifier, and for the "Non+" and "Fort+" configurations.

5.2.1 THREE LABEL CLASSIFIER

For the three label classifier, in Table 5.7, the best performances of `bert-base-multilingual-cased` are obtained with the fine-tuning on the three languages for precision on **Fort** class and on English only for recall.

Comparing the two models fine-tuned on the German dataset, `bert-base-german-cased` performs better on precision on **Fort** class, and `bert-base-multilingual-cased` better on recall. In overall the two results are comparable

Fine-tuning Language	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
FR	0,4754	0,4056	0,4377	0,7199	0,5101	0,5454
EN	0,3938	0,6224	0,4824	0,6308	0,5801	0,5626
DE	0,4651	0,5594	0,5079	0,7147	0,5712	0,5865
FR + EN	0,4727	0,5455	0,5065	0,6568	0,5755	0,5905
FR + EN + DE	0,4783	0,4615	0,4698	0,6691	0,5720	0,6031
DE	0,4740	0,5105	0,4916	0,6732	0,5751	0,5990

TABLE 5.1
Structuring questions detection. Three label classifier.
`bert-base-multilingual-cased` model, except last line: `bert-base-german-cased`.
Test on German.

Fine-tuning Language	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
FR	0,4722	0,2378	0,3163	0,7154	0,6115	0,644
EN	0,5094	0,1888	0,2755	0,7328	0,5893	0,6241
DE	0,5641	0,3077	0,3982	0,7632	0,6472	0,6863
FR + EN	0,5556	0,2797	0,3721	0,7582	0,6336	0,6730
FR + EN + DE	0,5490	0,3916	0,4571	0,7578	0,6868	0,7157
DE	0,5413	0,4126	0,4683	0,7545	0,6965	0,7211

TABLE 5.2
Structuring questions detection. “Non+” configuration.
`bert-base-multilingual-cased` model, except last line: `bert-base-german-cased`.
Test on German.

Fine-tuning Language	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
FR	0,7208	0,4531	0,5564	0,8341	0,7178	0,7605
EN	0,6255	0,6408	0,6331	0,7948	0,8013	0,798
DE	0,6681	0,6490	0,6584	0,8166	0,8084	0,8124
FR + EN	0,7010	0,5551	0,6196	0,8288	0,7658	0,793
FR + EN + DE	0,7273	0,5551	0,6296	0,842	0,7672	0,7987
DE	0,6933	0,6367	0,6638	0,8287	0,8043	0,8159

TABLE 5.3
Structuring questions detection. “Fort+” configuration.
`bert-base-multilingual-cased` model, except last line: `bert-base-german-cased`.
Test on German.

5.2.2 “NON+” & “FORT+”

In the “Non+” configuration, fine-tuning `bert-base-multilingual-cased` on German only gives the best precision on **Fort** class and the best recall is obtained with the three languages. For the two models fine-tuned on the German dataset, `bert-base-multilingual-cased` has a better precision and `bert-base-german-cased` a better recall.

In the "Fort+" configuration, fine-tuning `bert-base-multilingual-cased` on the three languages gives the best precision on **Fort** class and the best recall is obtained with German only. For the two models fine-tuned on the German dataset, `bert-base-german-cased` has a better precision and `bert-base-multilingual-cased` a better recall.

Comparing the two binary configurations shows once again that "Fort+" yields better results than "Non+" in every case.

5.3 INFLUENCE OF ORIGINAL LANGUAGE

The German dataset is heterogeneous in at least one point: it is translated from two different languages. Some sentences are originally in French and others in English. It is necessary to estimate the influence that could have the two initial languages on the performances and if one or another is pulling them up or down. This is done by taking again the predictions from the previous fine-tuning on French and English sets, but computing scores for each initial language separately. The results can be seen in Tables 5.4, respectively 5.5 and 5.6, for the three label classifier and respectively "Non+" and "Fort+" configurations.

Test Language	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
DE from FR	0,4468	0,5676	0,5000	0,6830	0,6008	0,6126
DE from EN	0,5070	0,5217	0,5143	0,5796	0,5297	0,5395

TABLE 5.4
Structuring questions detection. Three label classifier.
`bert-base-multilingual-cased` model. Fine-tuning on French and English datasets.
Test on German grouped by original language.

Test Language	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
DE from FR	0,5192	0,3649	0,4286	0,7433	0,6736	0,7017
DE from EN	0,6500	0,1884	0,2921	0,8014	0,5911	0,6325

TABLE 5.5
Structuring questions detection. "Non+" configuration.
`bert-base-multilingual-cased` model. Fine-tuning on French and English datasets.
Test on German grouped by original language.

Test Language	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
DE from FR	0,7131	0,6084	0,6566	0,8362	0,7913	0,8116
DE from EN	0,6806	0,4804	0,5632	0,8170	0,7298	0,7647

TABLE 5.6
Structuring questions detection. "Fort+" configuration.
`bert-base-multilingual-cased` model. Fine-tuning on French and English datasets.
Test on German grouped by original language.

For the three label classifier, for **Fort** class, the sentences originally in English have a better precision and the French one a better recall. The same goes for the "Non+" configuration and for "Fort+" the sentences initially in French have simultaneously a better precision and recall. Globally, no score is significantly higher than another, except for the French recalls of the binary classifiers that are however higher in a more marked way.

Test Language	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
FR	0,4651	0,5405	0,5000	0,7142	0,6047	0,6286
EN	0,5781	0,5362	0,5564	0,7337	0,5278	0,5462
DE	0,4800	0,5874	0,5283	0,7451	0,5851	0,6008

TABLE 5.7
Structuring questions detection. Three label classifier.
`bert-base-multilingual-cased` model. Fine-tuning on Multilingual dataset.

Test Language	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
FR	0,6522	0,4054	0,5000	0,8109	0,6971	0,7395
EN	0,7826	0,2609	0,3913	0,8698	0,6282	0,6836
DE	0,6610	0,2727	0,3861	0,8108	0,6325	0,6812

TABLE 5.8
Structuring questions detection. “Non+” configuration.
`bert-base-multilingual-cased` model. Fine-tuning on Multilingual dataset.

5.4 MULTILINGUAL FINE-TUNING

Now that the German dataset in Section 5.2 has shown that translated data could be used for fine-tuning and testing, the next step is to fine-tune `bert-base-multilingual-cased` on the twenty languages dataset from Figure 5.1. The results of these fine-tunings, i.e. the performances on the French, English and German test sets, can be seen in Tables 5.7, respectively 5.8 and 5.9, for the three label classifier and the “Non+” and “Fort+” configurations.

5.4.1 THREE LABEL CLASSIFIER

For the three label classifier, the English test set gets the best precision on class **Fort** and the best recall is obtained by the German test set. The recall of the three languages have the same order of magnitude, but the precision of the English set stands out from the two others. In comparison with the `bert-base-multilingual-cased` lines of Tables 4.9, 4.10 and 5.1, i.e. a bilingual fine-tuning (FR + EN) in the three label configuration, each language behaves differently. For French, the precision slightly decreases for an equivalent increase in recall. For English, precision and recall both decrease. For German, precision and recall both slightly increase.

The evolution of the performances when adding some languages in the fine-tuning is summarised in Figure 5.3. It is noticeable that for all testing language, there is no real influence of the number of fine-tuning languages, since all results are quite stable.

5.4.2 “NON+” & “FORT+”

For the “Non+” configuration, the best precision for **Fort** class is obtained by the English test set and the best recall is obtained by French. In comparison with the `bert-base-multilingual-cased` line of Tables 4.13, 4.14 and 5.2, i.e. a bilingual fine-tuning (FR + EN) in the “Non+” configuration, one can notice that the precision has significantly increased for French and English, but if the recall is practically unchanged for French, it has strongly decreased for English. For German, the same tendency appears to a lesser extent (slight increase in precision and decrease in recall).

For the “Fort+” configuration, the French test set obtains both the best precision and recall for **Fort** class.

Test Language	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
FR	0,8689	0,7413	0,8000	0,921	0,8647	0,8903
EN	0,8375	0,6569	0,7363	0,9032	0,8226	0,8574
DE	0,8232	0,6653	0,7359	0,8952	0,8255	0,8562

TABLE 5.9
Structuring questions detection. "Fort+" configuration.
`bert-base-multilingual-cased` model. Fine-tuning on Multilingual dataset.

In comparison with the `bert-base-multilingual-cased` line of Tables 4.17, 4.18 and 5.3, i.e. a bilingual fine-tuning (FR + EN) in the "Fort+" configuration, one can notice for the three languages a significant increase of precision and of recall (except for the German recall that is stable).

The evolution of the performances when adding some languages in the fine-tuning is summarised in Figures 5.4, respectively 5.5, for the "Non+", respectively "Fort+", configurations. In both cases, the addition of fine-tuning languages clearly increases the performances, in terms of precision and recall. In particular, the gap between the bilingual fine-tuning (FR + EN) and the multilingual with the twenty languages is significant.

5.5 ANALYSIS AND DISCUSSION

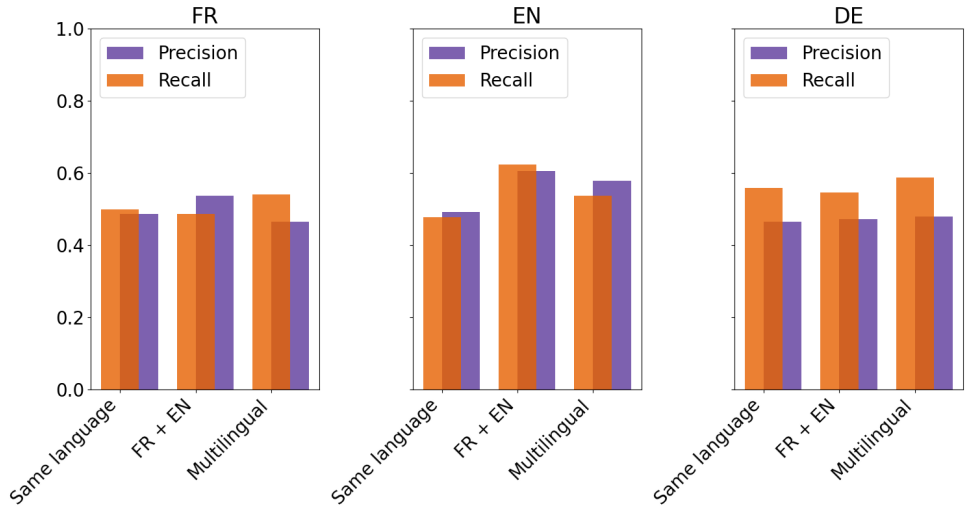


FIGURE 5.3
Evolution of precision and recall for **Fort** class, for the three label classifier, when adding fine-tuning languages.

The experiment proposed in this section gives a good insight of how the scope of every language of the Members of the EBU can be handled, in a context where the absence of experts in these languages is a major challenge. As shown in Figure 5.2, the translation seems to be trustable and coherent data, at least in comparison with the original speech-to-text transcriptions. However, this experiment should not be taken for more than it is, in other words a first draft and exploratory solution to this wide problematic, because it suffers from important drawbacks at this point. Indeed, the control and evaluation of the translation was performed on only one language and not on the whole dataset. Furthermore, the language chosen for evaluation, German, is not the furthest from the two originals, therefore a good translation could have been already expected and it is still plausible that problems with less close languages have not been spotted.

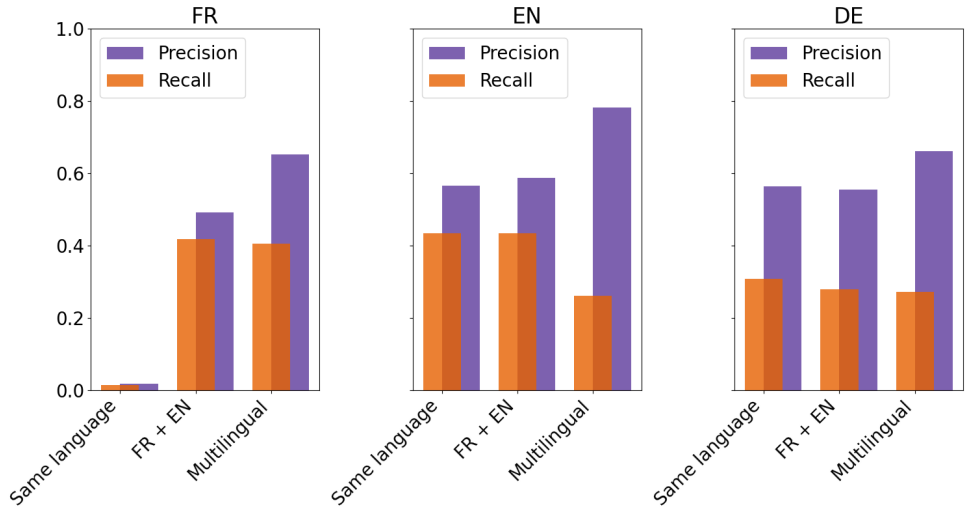


FIGURE 5.4
Evolution of precision and recall for **Fort** class, for the "Non+" configuration, when adding fine-tuning languages.

Another limitation of this method is that it is far from creating complete new data in these new eighteen languages. Certainly, it makes it possible to feed a multilingual model with the shape and structure of other languages, but fundamentally, what is being said is not new. It is probable that what is identified in the end to be the essence of a structuring sentence is redundant in fine-tuning data, and more than twice then. Finally, the data cannot be considered as representative of exact transcriptions of potential podcasts in these languages, for the reason that the practices must vary according to the country or the cultural context it is targeted for, in both form and content. Since it was automatically processed twice, first during speech-to-text process and then during translation, it is also likely that the final result differs from spoken language, in the sense that the translation tools might have sometimes a tendency to formalise some sentences, but will never spontaneously make them look more spoken.

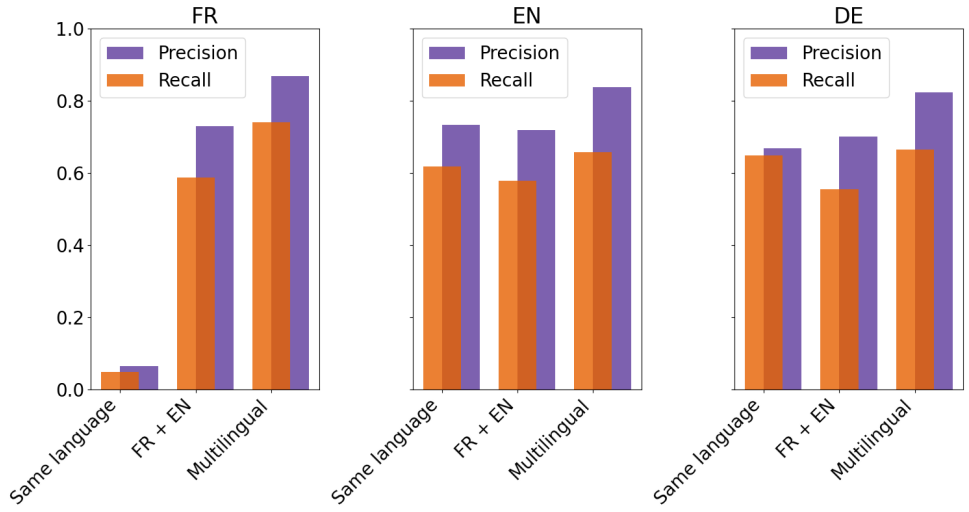


FIGURE 5.5
Evolution of precision and recall for **Fort** class, for the "Fort+" configuration, when adding fine-tuning languages.

This being said, the experiment shows some encouraging results. In particular, the quality of the German

translation allows making both fine-tuning and testing on these data. In Tables 5.4, 5.5 and 5.6, no real difference can be pointed out, despite the fact that the translation errors were mainly observed from originally French data than from English (see Figure 5.2b). The testing on the German data set shows in Tables 5.1, 5.2 and 5.3 a similar behaviour as in Figure 4.3, with a rather low-performance three label classifier and, regarding the binary classifiers, higher results with the "Fort+" configuration than with the "Non+". This is also confirmed by Tables 5.4, 5.5 and 5.6 where the same phenomenon is observed with both original languages. These elements suggest that the essence of what makes a sentence structuring for the podcast is largely language-independent.

However, the multilingual fine-tuning, as it is summed up in Figures 5.3, 5.4 and 5.5 shows in most of the cases a progression in comparison with the mono or bilingual case, in particular in the three tested languages in the "Fort+" configuration, where both precision and fort recall on **Fort** label increase.This indicates that the possibility of feeding models with data in many languages is still a matter of interset, in particular because this was observed on the three languages. That being said, it would be preferable to find experts on all languages to be able to verify if this really applies to a broader scope.

5.6 CONCLUSIONS

To summarise this chapter, the German tests are an indicator that the need of multilingual data for fine-tuning is not imperative to detect a structuring sentence, since their nature is partially independent of the language and the multilingual models seem to already be able to generalise sufficiently well, but the fact that the testing data is not considered as fully representative and that the multilingual fine-tuning tends to improve the performances, turn them into precious added value, if human resources are available for podcast selection, transcription and annotation in these languages.

Among the suggestions that can be made to explore more deeply the necessity of multilingual fine-tuning data, the first step might be to only create some original testing sets in the twenty languages, in other words a few original podcasts in those languages transcribed and annotated by experts. The smaller size of the test sets in comparison with a whole training set would not increase that much the human labour compared to the present work. In this perspective, one should be able to test the performances on real podcast transcription data of a `bert-base-multilingual-cased` model fine-tuned on the original French and English data. If the results are comparable to the German results in Tables 5.1, 5.2 and 5.3, this would motivate the creation of a multilingual fine-tuning set with original podcast transcriptions in these languages and annotated by their experts. In that perspective, the significance of *Meta radio* in the context of this thesis should be emphasized, as its availability to all EBU Members will provide at least qualitative feedback for assessing the relevance of the structure of the shows. This feedback will come from people who are intimately familiar with the specific needs and working methods of each language and country or region.

After this exploration of an opening to the scope of all languages of the EBU, it is time to see if the original method of Section 4 can be improved, in particular with the insertion of contextual information during fine-tuning and prediction.

CHAPTER 6

STRUCTURING WITH CONTEXT

So far, you have worked mainly with the text transcription of the podcasts. But there are much information that the podcast context provide that could also be used, right?

Back to the original non-translated data, the last idea of this thesis is to open the prediction of whether a sentence is structuring or not to some more context, since it is so far is only based on the content of the sentence itself. It has already been mentioned that podcasts are full of exploitable information, without talking about the possibility of introducing a window of text before and/or after the current sentence.

In this chapter, the head of the sentences are modified with the results of the diarisation and also with the timestamps in the podcast. This is also the opportunity to bind the two important points of this project: the host identification with the structuring questions detection. It seems obvious to connect these two in practice, since they are by nature linked by their definition, because the structuring sentences can only be pronounced by the host.

All the new experiments are performed in the "Fort+" configuration, since it is giving better results than the others in the previous sections.

6.1 WITH HOST DETECTION

The first modification made to the sentences is to add at their head the diarisation information, after the host discrimination within identified speakers has been performed. This means the BERT model is given the following as entry:

```
"SPEAKER_XX : And we will see some
images because even, you know, the
baguette has become, well, that's
not the figure, but like, for
example, bread makers, they have
had to stop working, stop their
ovens because, well, it was so
expensive."
```

with XX being the speaker id given by pyannote,

```
"HOST : I mean, that carbon
footprint that you're making
reference to, Shirley, can be
calculated for almost any kind
of territory, country, industry,
and also for each and every one of
us, if you look at it kind of more
in more detail."
```

if the speaker was predicted as being the host,

with the intuition that it should be able to identify that a **Fort** annotated sentence always starts with "HOST", or almost since there can be diarisation errors, but that in the meantime, beginning with "HOST" is not a sufficient condition to be **Fort**.

The fine-tuning is only performed on each language separately, but for the same eight models than in Section 4. The results are presented in Table 6.1 for French and Table 6.2 for English.

For French, the best precision on **Fort** class is obtained by xlm-roberta-large and the best recall by bert-base-multilingual-cased. The comparison of scores with Section 4 can be seen in Figure 6.2. It is however not really justified to compare the two, because of the unexplained poor performances of the French test set in Table 4.15.

For English, the best precision on **Fort** class is obtained by bert-large-cased and the best recall by bert-base-cased and bert-base-uncased simultaneously. The comparison of scores with Section 4 can be seen in Figure 6.3. It is clearly noticeable that precision and recall on **Fort** class increase for all models.

Model	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,6972	0,6923	0,6947	0,8323	0,8303	0,8313
bert-large-cased	0,7407	0,6993	0,7194	0,8546	0,8367	0,8453
bert-base-uncased	0,7686	0,6503	0,7045	0,8661	0,8148	0,838
bert-base-multilingual-cased	0,6800	0,7133	0,6962	0,8248	0,8389	0,8316
camembert-base	0,7111	0,6713	0,6906	0,8383	0,8212	0,8295
camembert-large	0,6944	0,6993	0,6969	0,8313	0,8334	0,8323
xlm-roberta-base	0,7333	0,6923	0,7122	0,8505	0,8328	0,8414
xlm-roberta-large	0,8103	0,6573	0,7259	0,8874	0,8205	0,8499

TABLE 6.1
Structuring questions detection. "Fort+" configuration.
Fine-tuning on French dataset. Test on French.
Entry: Speaker + Sentence

..

Model	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,8462	0,7549	0,7979	0,9119	0,8711	0,8902
bert-large-cased	0,8795	0,7157	0,7892	0,9269	0,8533	0,8859
bert-base-uncased	0,8750	0,7549	0,8105	0,9263	0,8725	0,8972
bert-base-multilingual-cased	0,8022	0,7157	0,7565	0,8881	0,8497	0,8677
camembert-base	0,8315	0,7255	0,7749	0,9032	0,8560	0,8778
camembert-large	0,8608	0,6667	0,7514	0,9153	0,8284	0,8656
xlm-roberta-base	0,7727	0,6667	0,7158	0,8712	0,8243	0,8458
xlm-roberta-large	0,7907	0,6667	0,7234	0,8802	0,8252	0,8500

TABLE 6.2
Structuring questions detection. "Fort+" configuration.
Fine-tuning on English dataset. Test on English.
Entry: Speaker + Sentence

6.2 WITH HOST DETECTION AND TIMESTAMPS

The second modification made to the head of the sentences is to add the diarisation information (again after host detection) and also the starting timestamp within the podcast. There is no particular reason to think there is a correlation between the absolute timestamp and the fact that a sentence is structuring, as it can be seen in Figure 6.1. In particular since there are many different show length in the dataset. The idea here is more to quickly and easily plug in some basic information about the podcast, with the intuition that it will make the text look more as extracted from a radio show rather than anywhere else, and see how it behaves.

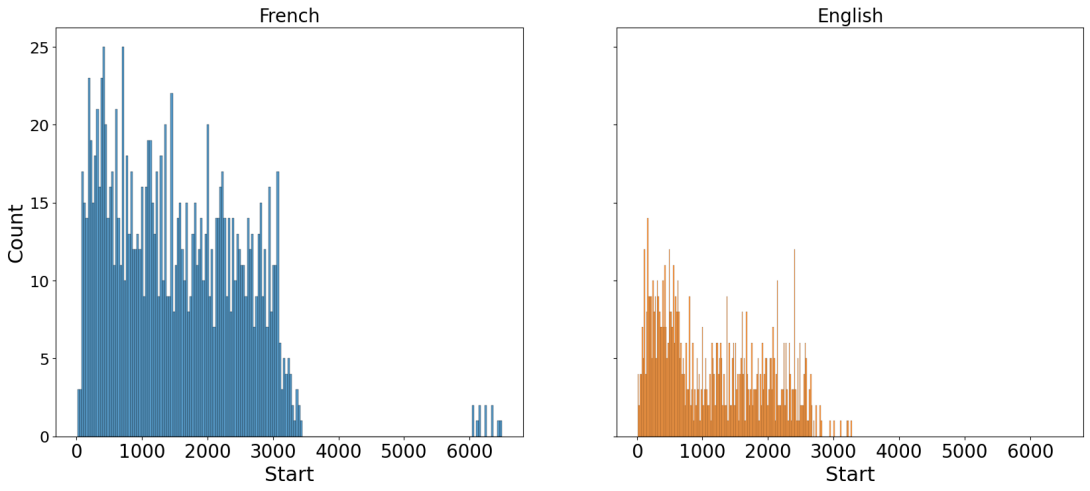


FIGURE 6.1
Distribution of **Faible** or **Fort** sentences within time.
This looks rather like noise, so no correlation appears between the two.

This means, if for example the sentence is pronounced at 37 min and 42 sec, BERT models are given the following in entry:

```
"SPEAKER_XX (00:37:42) :  And we
will see some images because even,
you know, the baguette has become,
well, that's not the figure, but
like, for example, bread makers,
they have had to stop working,
stop their ovens because, well,
it was so expensive."
```

with XX being the speaker id given by pyannote.

```
"HOST (00:37:42) :  I mean, that
carbon footprint that you're
making reference to, Shirley, can
be calculated for almost any kind
of territory, country, industry,
and also for each and every one of
us, if you look at it kind of more
in more detail."
```

if the speaker was predicted as being the host.

As previously, the fine-tuning is only performed on each language separately, for the same eight models. The results are shown in Table 6.3 for French and Table 6.4 for English.

For French, the best precision on **Fort** class is obtained by xlm-roberta-large and recall by xlm-roberta-base. The comparison of scores with less information can be seen in Figure 6.2. One can notice a tendency

that the timestamp information, in comparison with diarisation alone, increases slightly the precision with a rather stable recall.

For English, the best precision on **Fort** class is obtained by camembert-large and recall by bert-large-cased. As it already happened in previous experiments, xlm-roberta-large overfits and never predicts **Fort**. The comparison of scores with less information can be seen in Figure 6.3. Here the tendency is harder to identify, but there is a majority of cases where precision increases in comparison with cases where it is stable, decreases or is an overfitting case, which is more due to some model instability and makes it more relevant to be ignored.

Model	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,8482	0,6643	0,7451	0,9068	0,8259	0,8607
bert-large-cased	0,7760	0,6783	0,7239	0,8712	0,8288	0,8483
bert-base-uncased	0,8000	0,7273	0,7619	0,8857	0,8540	0,8690
bert-base-multilingual-cased	0,7021	0,6923	0,6972	0,8348	0,8306	0,8327
camembert-base	0,7153	0,6853	0,7000	0,8411	0,8282	0,8345
camembert-large	0,7609	0,7343	0,7473	0,8664	0,8549	0,8606
xlm-roberta-base	0,7267	0,7622	0,7440	0,8507	0,8660	0,8581
xlm-roberta-large	0,8900	0,6224	0,7325	0,9256	0,8071	0,8544

TABLE 6.3
Structuring questions detection. “Fort+” configuration.
Fine-tuning on French dataset. Test on French.
Entry: Speaker + Start + Sentence

Model	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,8444	0,7451	0,7917	0,9106	0,8662	0,8868
bert-large-cased	0,8021	0,7549	0,7778	0,8898	0,8689	0,879
bert-base-uncased	0,8261	0,7451	0,7835	0,9014	0,8653	0,8823
bert-base-multilingual-cased	0,8235	0,6863	0,7487	0,8975	0,8364	0,8638
camembert-base	0,8933	0,6569	0,7571	0,9312	0,8248	0,8689
camembert-large	0,9189	0,6667	0,7727	0,9445	0,8306	0,8774
xlm-roberta-base	0,8506	0,7255	0,7831	0,9128	0,8569	0,8823
xlm-roberta-large	0	0	0	0,4578	0,5000	0,4780

TABLE 6.4
Structuring questions detection. “Fort+” configuration.
Fine-tuning on English dataset. Test on English.
Entry: Speaker + Start + Sentence

6.3 ANALYSIS AND DISCUSSION

This chapter shows that contextual information is a fundamental need for the models to improve their performances or stabilise their functioning. It indicates that the richly detailed form of the podcasts is too precious to be put apart, the first being that it is a conversation between many speakers and that one has a special role: the host. It shows the importance of its identification during the process and that it is a logical step to bind the two predictions one after another. In that sense, even if Figure 6.1 shows there is no correlation between absolute time and structuring sentence, it is probable that the fact making the textual transcript look as if it was a screenplay of the show can help the models in the prediction.

However, why decide to add the token of the speaker, with "`HOST`" token for the host, instead of setting aside the sentences of the non-host speakers ? Indeed, their sentences have been defined as non-structuring in any case during annotation in Section 3.4.2, so it could be a legitimate choice to do the structuring prediction on host sentences only. There are two arguments that are in favour of the choice of keeping all sentences. The first one is that even if the host detection work well on diarisation, the latter is not perfect and some host sentences would therefore not be taken into account. By taking all sentences for structuring prediction, they are still given a chance to be predicted as **Fort**. The second is that this restriction to the host sentences for being structuring might be abandoned in future work in another context. None of the podcasts of the present work corresponds to that case, but there exists some where listeners are invited to take part in the show to ask questions. This can be the case in political debates, shows dedicated to consumers, or some more specific interactive podcasts.

After these initial tests, many ideas can be suggested to explore this path more deeply. For example, since Figures 6.2 and 6.3 show a global increase in performances, it would be interesting to test this addition of information on the three label classifier and on the "Non+" configuration, since it was established that "Fort+" configuration increase in precision might be due to a weaker **Fort** class, and it would be valuable to enhance the more difficult, but more interesting in practice, tasks.

On the sentence content enrichment point of view, many aspects can be considered. Firstly, one could modify not just the head of the sentence, but also its tail, with for example the ending timestamp of the sentence. Secondly, jingle information could be inserted, at head or tail, representing them by a special token or maybe by a description of them. The fact that the sentence appears at the beginning of a turn of speech could be also underlined with the insertion of a dash in French and quotes in English. This would however need to be adapted to each language. In addition to that, some audio analysis could be performed, like tone detection, and inserted in a written form in the sentences. Finally, one could also simply suggest, in particular after the insertion of all these elements, to insert a window of text. with more than just the sentence that is classified.

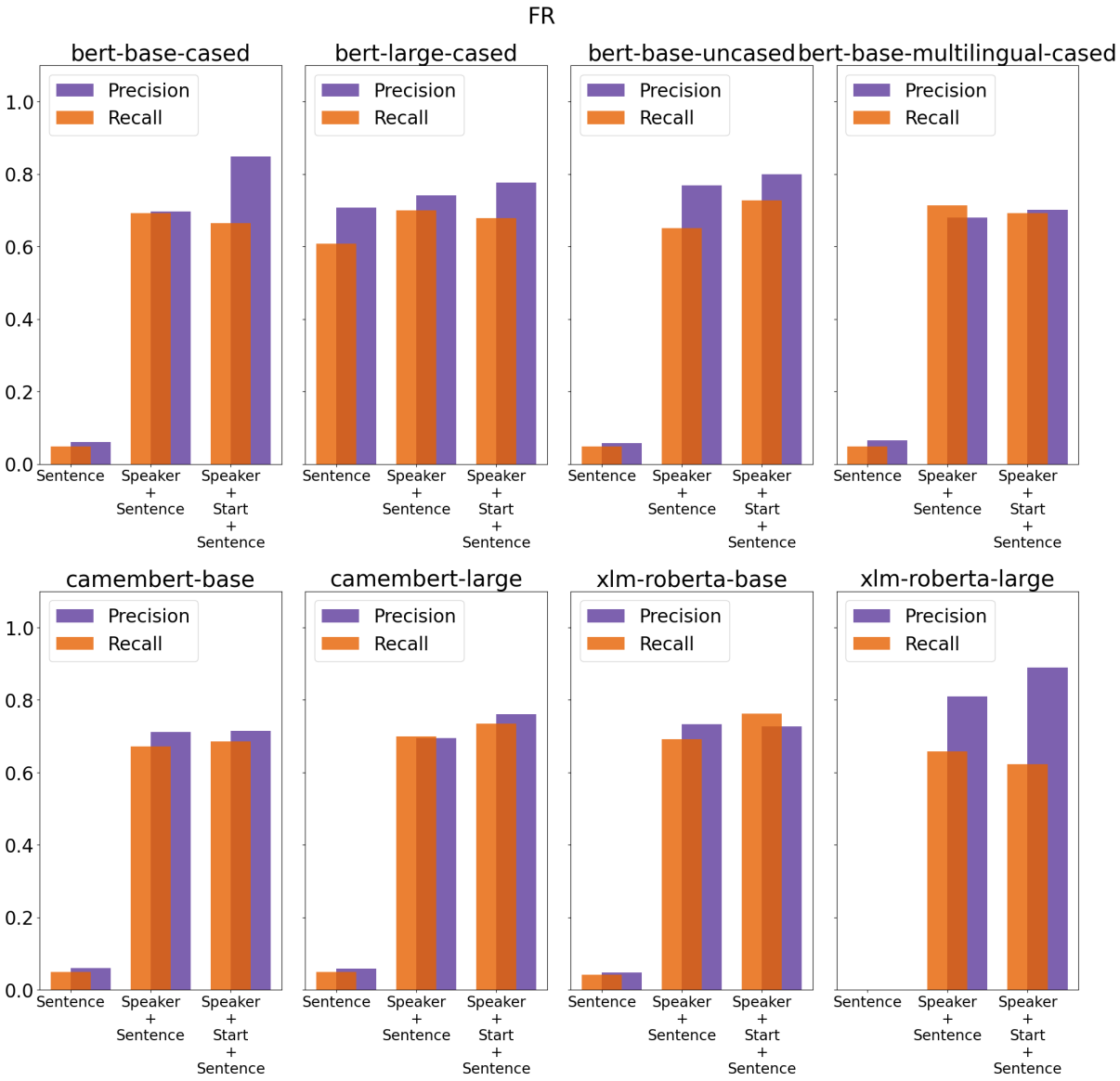


FIGURE 6.2
Performance comparison, for class **Fort**, between the levels of information given with the sentence, i.e. between Tables 4.15, 6.1 and 6.3.

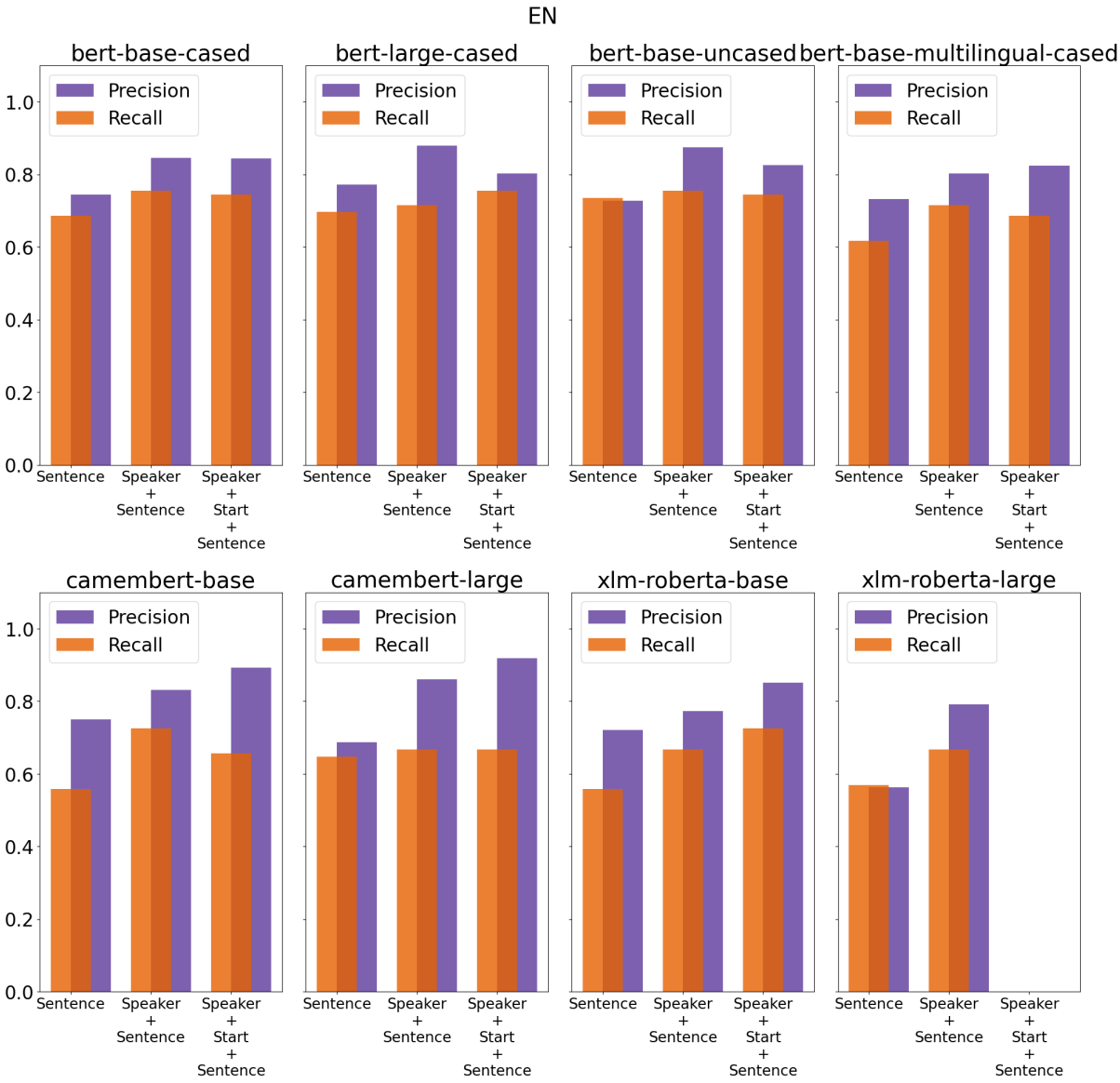


FIGURE 6.3
Performance comparison, for class **Fort**, between the levels of information given with the sentence, i.e. between Tables 4.16, 6.2 and 6.4.

CHAPTER 7

CONCLUSION

Time is always too short and we are already approaching the end of this show. One final word for our listeners? What sould they go home with after all this work?

7.1 RESULTS SUMMARY

This Master thesis demonstrates the benefits of fine-tuning BERT models to chapterise podcasts in a way specifically designed for that type of document. With adapted dataset and annotation, it results in satisfactory results and method simple to put into practice.

The first question raised in the introduction was the capability of the different BERT models. All of the models tested demonstrate potential for both host detection and structuring question detection tasks. However, the unstable behaviour of large models indicates a focus on base models may be more beneficial. One fortunate result, however, is that the multilingual models perform, in average, as well as the language-oriented models in that language. This finding suggests the feasibility of a tool that could be used by EBU Members, analysing their content directly in their own language.

Another question that arose was to compare French and English performances. If on the monolingual fine-tunings, English results are higher, the bilingual experiment shows that French can be raised to English level with a sufficient amount of data. This is also confirmed by the outcomes on the automatically translated German test dataset, that are comparable to the data in original language performances. However, if the translated data may be reliable on the point of view of the language, they might not be representative of the actual practice of podcast construction in the country where they are spoken. This is why the EBU *Meta Radio* tool is a great application perspective of this project, as it will allow collecting some qualitative feedback from the Members, which are the most qualified to evaluate this.

From the host detection perspective, many conclusions can be drawn. The most significant conclusion is that it is more effective to rely on diarisation first and then pinpoint from the list of speakers the one that is the host, as opposed to using a classifier that makes a prediction for each sentence independently, i.e. without taking in account the diarisation. Nonetheless, this project demonstrates that the host classifier can be sufficiently precise and simple enough to implement to be used for this detection task. Moreover, the initial rule-based method of this project, with all its weaknesses, is also a proof that pinpointing in the diarisation is an approach to be privileged. Even if the developed classifier can be considered a bit

specific, this is not totally problematic, as this tool is ultimately dedicated to a specific target: the EBU Members, who are likely to have similar needs.

Another point that has not been mentioned so far, is the legal implication of this method. Indeed, the host can be detected from a single and independent show, such that there is no need to either save its voice features or link the results with its identity. This saves a lengthy reflection on the data privacy protection set up, which can be an important issue for members who do not necessarily have the resources for this.

The structuring question classifier is by nature a more specific task and intended for shows that are precisely chosen to be conducted rigorously by the host. Even if this can be considered as the standard way of podcast production, there are many types of show on which the structuring question detection approach developed in this project cannot be applied. One can think for example about podcasts where there is more than one host, or shows where listeners are invited to actively participate and ask questions on air. These questions might be expressed differently than host do and therefore not be identified as structuring by the classifiers, even if they can be considered as so by the listeners. This is because the project was defined in this way at the beginning; considering more diverse types of podcasts raises interesting questions that could be considered in future work. For example, what might be the strategy for a show where no question is necessarily asked, like a sport event comment or a concert diffusion?

One lesson of this work is also that, as it is supposed at the start of this thesis, the speech-to-text and diarisation tools are reaching a degree of maturity that makes the transcription pipeline quite easy to implement nowadays. This development represents positive news, as the resources can be focused on other aspects, such as annotation of transcripts.

7.2 FUTURE WORK

Talking about future work, the developing ideas for this project in the future are numerous.

Firstly, one could think about extending the identification of the host to the labelling of each role in the podcast, such as determining who are the guests, at least maybe by simply assigning them some hierarchical importance. Linking each speaker to their exact identity might again be a matter of legal consideration, but being able to determine their importance, whether there is a main guest or not, might be useful information to extract. In that perspective, the presentation of the guests by the host might sound as a starting point.

Secondly, one could also think about joining the two classifying tasks of this project into one multitask classifier. The two being by nature correlated, it might be possible that the two tasks can help each other.

Furthermore, it could also argue that the classifiers are missing a crucial element in the detection of structuring questions: the answer. It may be worth thinking about political shows, where politicians might tend to avoid answering the specific question asked, or about very prolix guests, who might be prone to digression. In a word, an interesting answer does not necessarily appear after a structuring question and, vice versa, very informative content can be found after an innocent intervention.

Another suggestion that could be made is to work more closely to with the production of podcasts themselves. Undoubtedly, working from the final audio file, which is the final product of the podcast, seems to rely on poorly detailed data. For richer information, the multichannel audio from the control room could be saved, knowing that the host probably always speaks in the same microphone. This could also be used to treat apart the prerecorded elements.

From this perspective, the written notes of the producers could also be used. There is good reason to estimate that the host probably writes down the main skeleton he/she wants to give to the show. Therefore, one could try to conciliate these notes with the transcription. However, such an approach would inevitably

have to deal with noisy notes, that might differ a lot from a host to another and require a well-organised logistic to collect and archive these notes properly with the audio.

Given the industrial context of this work, a future direction could also be to further adapt approaches to specific application needs. For example, after having processed a show, one could limit the number of structuring questions to some maximum and filter out false positives manually. The interest for a public service media is more to develop an assisting tool for a social media manager or for a documentalist. In practice, a human treatment should always follow, treating the output of classifiers as suggestions.

It could be relevant to also question the capacity of the developed tool to process archives of the EBU Members. Probably that good performances can be expected for shows that are similar to the dataset of this project, but the quality of the speech-to-text and diarisation services may decline and the hypothesis that they required no evaluation reconsidered. These tools may be more sensitive to old accents that can be found in archives or have more difficulties with transcribing proper noun of personalities or entities from the past.

One has also to underline that the implementation of such classifier in practice, might not replace other chapterisation methods, based on topics, or audio analysis with jingles. This work should rather be seen as an additional filter that could be cross-checked with other options.

Finally, an essential suggestion that can be raised is the use of generative Large Language Models (LLM). This point in fact, could seem to be missing in this Master thesis, as they recently became very popular, with the emergence of ChatGPT[1] in the first place and all the models that followed. It was chosen not to adopt them for two main reasons. The first is that public service media have a fundamental need for independence, that encourages to give priority to open-source models as much as possible, but they had not yet reached the same standard as the proprietary solutions at the beginning of the project. The second is that, even if the concession had been made to use proprietary models, there was no existing one that combined affordability and high performance for the specific task of chapterisation. However, this particular field having developed drastically fast, some open-source models, as Mixtral[2], appear as a great opportunity at the end of the project.

Regarding the usage of generative LLM, the work of this project could offer valuable insights that can serve as a guide for enhancing prompts with podcast specificities. For example, in addition to the transcript of the podcast and the instruction of proposing chapters in it, the LLM could be given the list of structuring questions asked by the host, or simply the transcript under the form of a screenplay, detailing the role of each speaker. In other word, there is a great opportunity to perform Chain-of-Thought prompting to guide the LLM through a chapterisation task. Another simple application would be to use LLMs as generators for chapter titles that appear between two structuring questions. In that sense, they could be useful to reformulate the questions in an efficient, clear and unambiguous manner.

As a final word, it is important to recall that this project aims to provide a fresh listening experience for the audience. However, there are probably as many ways to listen to a podcast as there are listeners. This project should be, in this context, seen as one particular filter that can be integrated in an environment where several alternatives are proposed, to allow the user to personalise its way of navigating in this ocean of audio content at its ease and its "*vaste appétit*". This thesis calls for the development of additional methods to enrich audio content, specifically through the creation of more HOST: Highly Organised Show Transcriptions.

[1] `chat.openai.com`
[2] `https://mistral.ai/fr/news/mixtral-of-experts/`

CHAPTER 8

APPENDIX

8.1 CROSS LANGUAGE TESTS

Model	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,2526	0,3478	0,2927	0,4660	0,4591	0,4532
bert-large-cased	0,1170	0,7391	0,2020	0,4309	0,5252	0,3920
bert-base-uncased	0,1547	0,4058	0,2240	0,3957	0,4996	0,3785
bert-base-multilingual-cased	0,4384	0,4638	0,4507	0,5478	0,5261	0,5328
camembert-base	0,5000	0,1159	0,1882	0,5821	0,4201	0,4515
camembert-large	0	0	0	0,3052	0,3333	0,3186
xlm-roberta-base	0,4915	0,4203	0,4531	0,6052	0,4968	0,5211
xlm-roberta-large	0	0	0	0,3053	0,3318	0,3180

TABLE 8.1
Structuring questions detection. Three label classifier.
Fine-tuning on French dataset. Test on English

Model	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,3714	0,1757	0,2385	0,5493	0,5768	0,5353
bert-large-cased	0,3636	0,0541	0,0941	0,5420	0,5310	0,4838
bert-base-uncased	0	0	0	0,4038	0,5572	0,4364
bert-base-multilingual-cased	0,3488	0,4054	0,3750	0,5755	0,5665	0,5687
camembert-base	0,3714	0,3514	0,3611	0,7293	0,4798	0,5022
camembert-large	0,3947	0,2027	0,2679	0,6961	0,5030	0,5591
xlm-roberta-base	0,3072	0,6351	0,4141	0,4228	0,5298	0,4573
xlm-roberta-large	0	0	0	0,3014	0,3333	0,3166

TABLE 8.2
Structuring questions detection. Three label classifier.
Fine-tuning on English dataset. Test on French

Model	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,4000	0,2319	0,2936	0,6773	0,6054	0,6301
bert-large-cased	0,1987	0,4493	0,2756	0,5813	0,6698	0,6006
bert-base-uncased	0,3846	0,3623	0,3731	0,6731	0,6636	0,6682
bert-base-multilingual-cased	0,4324	0,2319	0,3019	0,6936	0,6067	0,6349
camembert-base	0,8000	0,0580	0,1081	0,8730	0,5285	0,54000
camembert-large	0	0	0	0,4714	0,5000	0,4853
xlm-roberta-base	0,4898	0,3478	0,4068	0,7255	0,6629	0,6882
xlm-roberta-large	0	0	0	0,4714	0,5000	0,4853

TABLE 8.3
Structuring questions detection. “Non+” configuration.
Fine-tuning on French dataset. Test on English

Model	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	1,0000	0,0135	0,0267	0,9756	0,5068	0,5008
bert-large-cased	0	0	0	0,4753	0,5000	0,4873
bert-base-uncased	0,8000	0,0541	0,1013	0,8765	0,5267	0,5384
bert-base-multilingual-cased	0,3966	0,3108	0,3485	0,6805	0,6431	0,6592
camembert-base	0,3810	0,4324	0,4051	0,6756	0,6979	0,6859
camembert-large	0	0	0	0,4753	0,5000	0,4873
xlm-roberta-base	0,4889	0,2973	0,3697	0,7265	0,6406	0,6718
xlm-roberta-large	0	0	0	0,4753	0,5000	0,4873

TABLE 8.4
Structuring questions detection. “Non+” configuration.
Fine-tuning on English dataset. Test on French

Model	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,4417	0,5196	0,4775	0,6983	0,7295	0,7123
bert-large-cased	0,5447	0,6569	0,5956	0,7562	0,8031	0,7770
bert-base-uncased	0,3772	0,6176	0,4684	0,6699	0,7618	0,7009
bert-base-multilingual-cased	0,6104	0,4608	0,5251	0,7809	0,7168	0,7436
camembert-base	0,8333	0,2941	0,4348	0,8859	0,6443	0,7003
camembert-large	0,8261	0,3725	0,5135	0,8855	0,6827	0,7409
xlm-roberta-base	0,6835	0,5294	0,5967	0,8205	0,7534	0,7820
xlm-roberta-large	0	0	0	0,4578	0,5000	0,4780

TABLE 8.5
Structuring questions detection. “Fort+” configuration.
Fine-tuning on French dataset. Test on English

Model	Fort precision	Fort recall	Fort f1-score	macro avg precision	macro avg recall	macro avg f1-score
bert-base-cased	0,6268	0,6224	0,6246	0,7934	0,7916	0,7925
bert-large-cased	0,7216	0,4895	0,5833	0,8347	0,7348	0,7735
bert-base-uncased	0,5655	0,6643	0,6109	0,7647	0,8052	0,7829
bert-base-multilingual-cased	0,6429	0,5035	0,5647	0,7958	0,7370	0,7621
camembert-base	0,4751	0,7343	0,5769	0,7226	0,8242	0,7591
camembert-large	0,6190	0,6364	0,6276	0,7902	0,7975	0,7938
xlm-roberta-base	0,6176	0,5874	0,6022	0,7871	0,7745	0,7806
xlm-roberta-large	0,4079	0,4336	0,4203	0,6738	0,6835	0,6784

TABLE 8.6
Structuring questions detection. “Fort+” configuration.
Fine-tuning on English dataset. Test on French

BIBLIOGRAPHY

Devlin, Jacob et al. (11th Oct. 2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805[cs]. URL: http://arxiv.org/abs/1810.04805 (visited on 19th Oct. 2023).

Hearst, Marti A. (1997). 'Text Tiling: Segmenting Text into Multi-paragraph Subtopic Passages'. In: *Computational Linguistics* 23.1, pp. 33–64. URL: https://aclanthology.org/J97-1003 (visited on 16th Oct. 2023).

Yoong, Siang Yun, Yao-Chung Fan and Fang-Yie Leu (2021). 'On Text Tiling for Documents: A Neural-Network Approach'. In: *Advances on Broad-Band Wireless Computing, Communication and Applications*. Ed. by Leonard Barolli et al. Lecture Notes in Networks and Systems. Cham: Springer International Publishing, pp. 265–274. ISBN: 9783030611088. DOI: 10.1007/978-3-030-61108-8_26.

Solbiati, Alessandro et al. (24th June 2021). *Unsupervised Topic Segmentation of Meetings with BERT Embeddings*. arXiv: 2106.12978[cs]. URL: http://arxiv.org/abs/2106.12978 (visited on 13th Oct. 2023).

Heiselberg, Lene and Iben Have (4th Apr. 2023). 'Host Qualities: Conceptualising Listeners' Expectations for Podcast Hosts'. In: *Journalism Studies* 24.5, pp. 631–649. ISSN: 1461-670X, 1469-9699. DOI: 10.1080/1461670X.2023.2178245. URL: https://www.tandfonline.com/doi/full/10.1080/1461670X.2023.2178245 (visited on 30th Nov. 2023).

Charlet, Delphine (Mar. 2010). 'Model-free anchor speaker turn detection for automatic chapter generation in broadcast news'. In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. ISSN: 2379-190X, pp. 4966–4969. DOI: 10.1109/ICASSP.2010.5495090. URL: https://ieeexplore.ieee.org/abstract/document/5495090 (visited on 1st Nov. 2023).

Yella, Sree Harsha, Vasudeva Varma and Kishore Prahallad (Dec. 2010). 'Significance of anchor speaker segments for constructing extractive audio summaries of broadcast news'. In: *2010 IEEE Spoken Language Technology Workshop*. 2010 IEEE Spoken Language Technology Workshop, pp. 13–18. DOI: 10.1109/SLT.2010.5700815. URL: https://ieeexplore.ieee.org/abstract/document/5700815 (visited on 1st Nov. 2023).

Louradour, Jérôme (2023). *whisper-timestamped*. https://github.com/linto-ai/whisper-timestamped.

Radford, Alec et al. (2022). 'Robust speech recognition via large-scale weak supervision'. In: *arXiv preprint arXiv:2212.04356*.

Plaquet, Alexis and Hervé Bredin (2023). 'Powerset multi-class cross entropy loss for neural speaker diarization'. In: *Proc. INTERSPEECH 2023*.