ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

DIGITAL HUMANITIES LABORATORY

MASTER THESIS

Where Did the News come from?

Detection of News Agency Releases in Historical Newspapers

LEA MARXEN

SUPERVISORS: Maud Ehrmann (DHLAB, EPFL), Emanuela Boros (DHLAB, EPFL), Marten Düring (C2DH, uni.lu)
PROFESSORS: Frédéric Kaplan (DHLAB, EPFL), Bastian Leibe (Visual Computing Institute, RWTH)

18 August 2023

Abstract

Since their beginnings in the 1830s and 1840s, news agencies have played an important role in the national and international news market, aiming to deliver news as fast and as reliable as possible. While we know that newspapers have been using agency content for a long time to produce their stories, the amount to which the agencies shape our news often remains unclear. Although researchers have already addressed this question, recently by using computational methods to assess the influence of news agencies at present, large-scale studies on the role of news agencies in the past continue to be rare.

This thesis aims to bridge this gap by detecting news agencies in a large corpus of Swiss and Luxembourgish newspaper articles (the *impresso* corpus) for the years 1840-2000 using deep learning methods. For this, we first build and annotate a multilingual dataset with news agency mentions, which we then use to train and evaluate several BERT-based agency detection and classification models. Based on these experiments, we choose two models (for French and German) for the inference on the *impresso* corpus. Results show that ca. 10% of the articles explicitly reference news agencies, with the greatest share of agency content after 1940, although systematic citation of agencies already started slowly in the 1910s. Differences in the usage of agency content across time, countries and languages as well as between newspapers reveal a complex network of news flows, whose exploration provides many opportunities for future work.

Acknowledgements

First and foremost, I would like to say an enormous thank you to my supervisors, whose guidance and help was invaluable for the project. With their kindness, their sense of purpose and their readiness to support me as well as their trust in my work, they created a working environment I was happy to be a part of every day. I also want to thank Maud Ehrmann, Emanuela Boros and Marten Düring specifically for their participation in the annotation campaign, which allowed me to distribute the workload of manually annotating 1,600 articles on four pairs of shoulders instead of one.

A big thanks goes out to all members of the DHLAB, who gave me a warm welcome and let me be part of their research world for one semester. Thank you for the conversations, both the serious and the nonsensical ones, and for accepting me as a short-term, but full member of your lab.

Last but not least, I would like to thank my friends and family for their love, their belief in me and their encouragement to follow my own path, even if it has often been the road less travelled by.

Contents

List of Abbreviations

This is a list of the main abbreviations used in this report. For a full list of all acronyms and abbreviations of newspapers and news agencies, see Appendix A and B.1.

Abbreviation	Definition
AFP	Agence France Presse
ANP	Algemeen Nederlands Persbureau
ANSA	Agenzia Nationale Stampa Associata
AP	Associated Press
ATP	Association of Tennis Professionals
ATS	Agence Télégraphique Suisse
BERT	Bidirectional Encoder Representations from Transformers
BTA	Bulgarska Telegrafitscheka Agentzia
CNN	Convolutional Neural Network
CRF	Conditional Random Fields
CTK	Czechoslavenska Tiskova Kancelar
DAPD	Deutscher Auslands-Depeschendienst
DDP	Deutscher Depeschendienst
DHLAB	Digital Humanities Laboratory
DL	Deep Learning
DPA	Deutsche Presseagentur
FN	False Negative
FP	False Positive
GPT	Generative Pre-trained Transformers
HIPE	Identifying Historical People, Places and Other Entities
IAA	Inter-Annotator Agreement
Kipa	Katholische Internationale Presseagentur
LED	Levenshtein Distance
LSTM	Long Short-Term Memory
ML	Machine Learning
MLM	Masked Language Model
NE	Named Entity
NER	Named Entity Recognition
NLP	Natural Language Processing
NSP	Next Sentence Prediction
OCR	Optical Character Recognition
OLR	Optical Layout Recognition
PAP	Polska Agencja Prasowa

List of Abbreviations (cont.)

Abbreviation	Definition
RNN	Recurrent Neural Network
SDA	Schweizerische Depeschenagentur
Tanjug	Telegrafska Agencija nova Jugoslavija
TASS	Telegrafnoie Agenstvo sovietskavo Soyusa
TP	True Positive
TT	Tidningarnas Telegrambyra
UNESCO	United Nations Educational, Scientific and Cultural Organization
UP	United Press
UPI	United Press International

1 Introduction

1.1 Context and Motivation

At the beginning of the 19th century, news from abroad was both slow to arrive and costly to acquire. Most of the time, newspapers needed to rely on reports from foreign newspapers or on private dispatches to third parties, re-transmitted to the newspaper editors, to be published the next day. Thus, references to newspapers' sources sometimes read like a who-is-who of the European news world and diplomatic landscape of that time[1]. Only those newspapers fortunate enough to have the means could maintain a few foreign correspondents who provided them with exclusive news stories from abroad.

The first news agencies were born out of this situation in the 1830s and 1840s. Founded privately or by coalitions of newspapers to save costs, they aimed at delivering news as fast, circumspect and accurate as possible (M. B. Palmer 2019, p. 17). Soon, they became key actors in the news world, and continue to function as such until today. However, their involvement in the production of journalistic content was (and is) not always visible or known, as they generally provided their services to newspapers and governments rather than directly to the public readership, thus remaining "invisible wholesalers" (Shrivastava 2007, p. 1) in the background.

Questions that naturally arise are: Can we still trace the influence of news agencies on the media, i.e. how invisible were they really? Were they systematically credited by most of the newspapers or not? How and to what extent did journalists rely on agency content to produce their stories? Was agency content used simply verbatim (copy and paste) or with rephrasing?

This Master's thesis aims to support the study of the role of news agencies in the media ecosystem and the use of their dispatches by the press over time, focusing on a corpus of digitised Swiss and Luxembourgish newspapers. It is embedded in the project *impresso – Media Monitoring of the Past*[2], which strives to develop new digital approaches to explore the media of the past across time, languages and countries (Ehrmann, Romanello, Clematide et al. 2020).

Concretely, the *impresso* corpus consists of 600,919 issues coming from 76 Swiss and Luxembourgish newspaper titles. Originally in paper form, they underwent a digitization process, including the scanning of each page, followed by optical layout recognition (OLR) and optical character recognition (OCR) processes to segment the page into different sections and turn the image of the text into digital text (see Figure 1.1). The *impresso* project collected the digitised versions from several sources, standardised them and enriched the text and metadata through various text mining algorithms. The enrichment allows researchers to use diverse approaches to investigate the corpus. Among them are topic modelling and named entity processing algorithms, which make it possible to identify newspaper articles based on the topics, persons and locations they contain. The computation of text reuse clusters based on text similarity

[1] See, for example, this article of the *Gazette de Lausanne* from Friday, November 22, 1822 (*impresso*, retrieved on 22.07.23)

[2] *impresso*. Media Monitoring of the Past. Supported by the Swiss National Science Foundation under grant CR-SII5_173719, 2017-2020. `https://impresso-project.ch`.; a second project phase will start September (2023-2027).

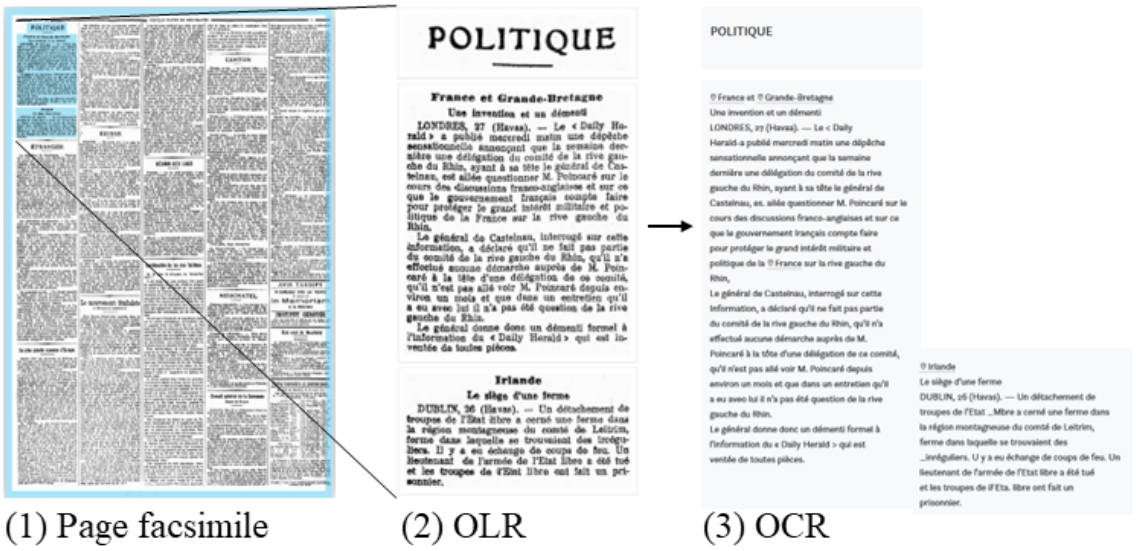


FIGURE 1.1
The digitization pipeline for a newspaper article: From (1) the newspaper facsimile, a scan of the whole page to (2) Optical Layout Recognition of article boundaries to (3) the transformation into text by Optical Character Recognition.

enables easier examination and comparison of how text was reused and altered throughout different newspapers. The *impresso* project also provides word embeddings[3], which can be used to search for similar words and concepts during the investigation of the corpus, but also to enhance machine learning algorithms trained on the *impresso* data. As a last step, a web interface[4] was designed to make the digitised data easily searchable for anyone interested in historical research.

The *impresso* corpus features newspapers in several languages (French, German and Luxembourgish). Together with the scale of data and its different enrichments, it provides new opportunities to conduct historical media research. This Master thesis takes part in the *impresso* project in various ways, on the one hand using the developed functionalities to explore the newspaper articles to find answers to the questions mentioned above, and on the other contributing to the project by adding a (news agency) layer to the metadata of the corpus.

1.2 Objective

The main objective of this thesis is the detection of news agency content in the *impresso* corpus. As Figure 1.2 shows, journalists are given different options for how to present an agency release in a newspaper. They can simply copy-paste the text, thus publishing the news agency dispatch in its original form, or they can alter the text to various degrees. Another choice is to credit or cite the news agency, as opposed to passing the content as one's own (Boyd-Barrett and M. Palmer 1981, p. 19ff). Hence, the evidence of news agency content within a particular newspaper article is not often easily found.

Because the *impresso* corpus consists solely of newspaper articles, not news agency releases, this "invisibility" of the news sources poses a major challenge for this project: If a text from an agency is not credited and is heavily edited, it will be nearly impossible to identify it as having agency content. Yet, this rarely seems to be the case (Boyd-Barrett and M. Palmer 1981, p. 19f). An article which has been credited as

[3] fastText word embeddings, `https://fasttext.cc`
[4] `https://impresso-project.ch/app/`

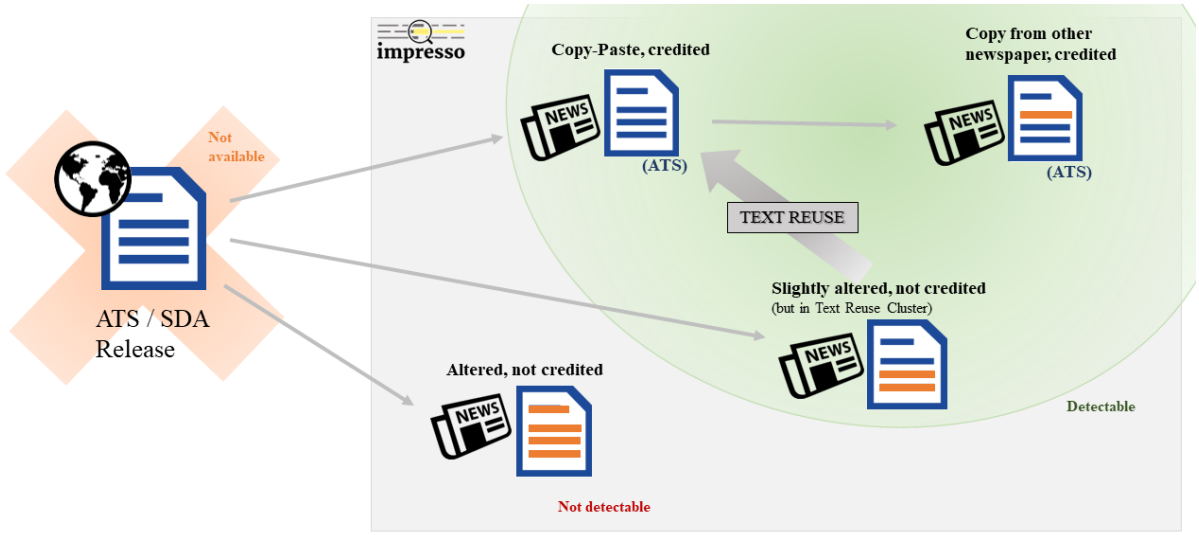
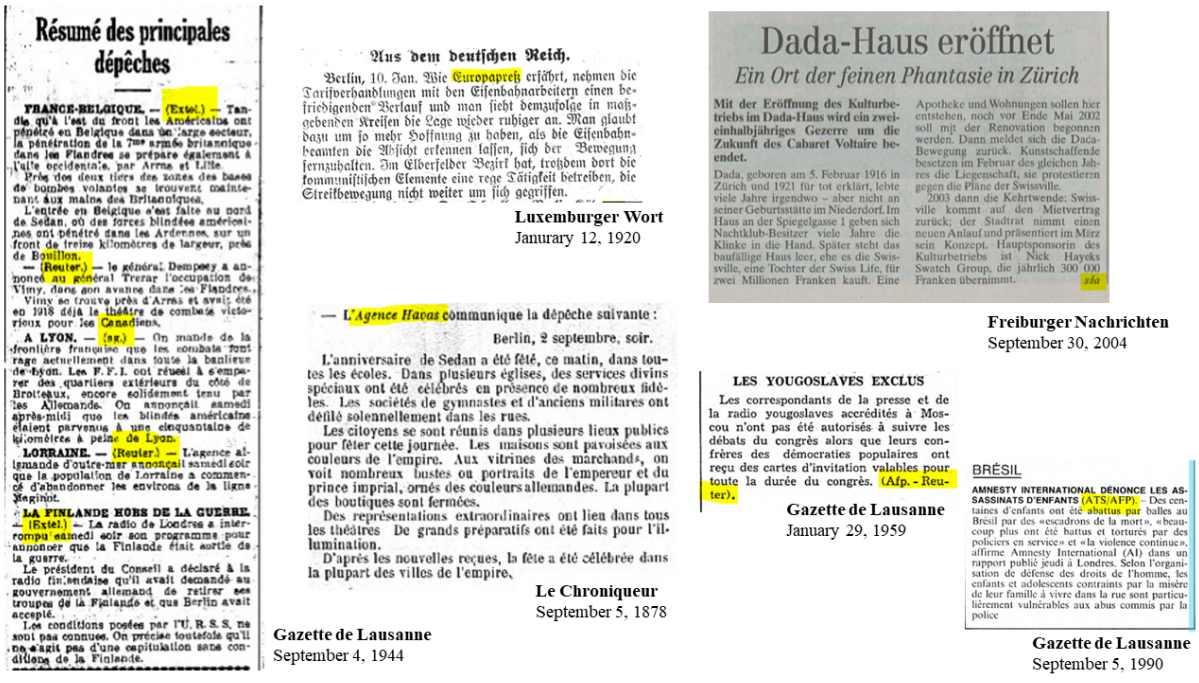


FIGURE 1.2

Illustration of the different ways a news agency release can be used in newspapers. In the absence of original agency releases (left), only articles that explicitly credit the news agency they are based on or those whose content is very similar to a credited agency release can be easily identified (green circle). Articles without explicit agency attribution and heavily altered content are almost impossible to detect (grey area).



Résumé des principales dépêches

FRANCE-BELGIQUE. — (Extel.) — Tandis qu'à l'est du front les Américains ont pénétré en Belgique dans un large secteur, la pénétration de la 7me armée britannique dans les Flandres se prépare également à l'aile occidentale, par Arras et Lille.

Près des deux tiers des zones des bases de bombes volantes se trouvent maintenant aux mains des Britanniques.

L'entrée en Belgique s'est faite au nord de Sedan, où des forces blindées américaines ont pénétré dans les Ardennes, sur un front de treize kilomètres de largeur, près de Bouillon.

— (Reuter.) — Le général Dempsey a annoncé au général Crerar l'occupation de Vimy, dans son avance dans les Flandres.

Vimy se trouve près d'Arras et avait été en 1918 déjà le théâtre de combats victorieux pour les Canadiens.

A LYON. — (ag.) — On mande de la frontière française que les combats font rage actuellement dans toute la banlieue de Lyon. Les F. F. I. ont réussi à s'emparer des quartiers extérieurs du côté de Brotteaux, encore solidement tenu par les Allemands. On annonçait samedi après-midi que les blindés américains étaient parvenus à une cinquantaine de kilomètres à peine de Lyon.

LORRAINE. — (Reuter.) — L'agence allemande d'outre-mer annonçait samedi soir que la population de Lorraine a commencé d'abandonner les environs de la ligne Maginot.

LA FINLANDE HORS DE LA GUERRE. — (Extel.) — La radio de Londres a interrompu samedi soir son programme pour annoncer que la Finlande était sortie de la guerre.

Le président du Conseil a déclaré à la radio finlandaise qu'il avait demandé au gouvernement allemand de retirer ses troupes de la Finlande et que Berlin avait accepté.

Les conditions posées par l'U. R. S. S. ne sont pas connues. On précise toutefois qu'il ne s'agit pas d'une capitulation sans conditions de la Finlande.

Gazette de Lausanne
September 4, 1944

Aus dem deutschen Reich.

Berlin, 10. Jan. Wie Europapreß erfährt, nehmen die Tarifverhandlungen mit den Eisenbahnarbeitern einen befriedigenden Verlauf und man sieht demzufolge in maßgebenden Kreisen die Lage wieder ruhiger an. Man glaubt dazu um so mehr Hoffnung zu haben, als die Eisenbahnbeamten die Absicht erkennen lassen, sich der Bewegung fernzuhalten. Im Elberfelder Bezirk hat, trotzdem dort die kommunistischen Elemente eine rege Tätigkeit betreiben, die Streikbewegung nicht weiter um sich gegriffen.

Luxemburger Wort
Janurary 12, 1920

— L'*Agence Havas* communique la dépêche suivante :

Berlin, 2 septembre, soir.

L'anniversaire de Sedan a été fêté, ce matin, dans toutes les écoles. Dans plusieurs églises, des services divins spéciaux ont été célébrés en présence de nombreux fidèles. Les sociétés de gymnastes et d'anciens militaires ont défilé solennellement dans les rues.

Les citoyens se sont réunis dans plusieurs lieux publics pour fêter cette journée. Les maisons sont pavoisées aux couleurs de l'empire. Aux vitrines des marchands, on voit nombreux bustes ou portraits de l'empereur et du prince imprial, ornés des couleurs allemandes. La plupart des boutiques sont fermées.

Des représentations extraordinaires ont lieu dans tous les théâtres. De grands préparatifs ont été faits pour l'illumination.

D'après les nouvelles reçues, la fête a été célébrée dans la plupart des villes de l'empire.

Le Chroniqueur
September 5, 1878

Dada-Haus eröffnet

Ein Ort der feinen Phantasie in Zürich

Mit der Eröffnung des Kulturbetriebs im Dada-Haus wird ein zweieinhalbjähriges Gezerre um die Zukunft des Cabaret Voltaire beendet.

Dada, geboren am 5. Februar 1916 in Zürich und 1921 für tot erklärt, lebte viele Jahre irgendwo – aber nicht an seiner Geburtsstätte im Niederdorf. Im Haus an der Spiegelgasse 1 geben sich Nachtklub-Besitzer viele Jahre die Klinke in die Hand. Später steht das baufällige Haus leer, ehe es die Swissville, eine Tochter der Swiss Life, für zwei Millionen Franken kauft. Eine Apotheke und Wohnungen sollen hier entstehen, noch vor Ende Mai 2002 soll mit der Renovation begonnen werden. Dann meldet sich die Dada-Bewegung zurück. Kunstschaffende besetzen im Februar des gleichen Jahres die Liegenschaft, sie protestieren gegen die Pläne der Swissville.

2003 dann die Kehrtwende: Swissville kommt auf den Mietvertrag zurück; der Stadtrat nimmt einen neuen Anlauf und präsentiert im März sein Konzept. Hauptsponsorin des Kulturbetriebs ist Nick Hayeks Swatch Group, die jährlich 300 000 Franken übernimmt. *sda*

Freiburger Nachrichten
September 30, 2004

LES YOUGOSLAVES EXCLUS

Les correspondants de la presse et de la radio yougoslaves accrédités à Moscou n'ont pas été autorisés à suivre les débats du congrès alors que leurs confrères des démocraties populaires ont reçu des cartes d'invitation valables pour toute la durée du congrès. (Afp. - Reuter).

Gazette de Lausanne
January 29, 1959

BRÉSIL

AMNESTY INTERNATIONAL DÉNONCE LES ASSASSINATS D'ENFANTS (ATS/AFP). – Des centaines d'enfants ont été abattus par balles au Brésil par des «escadrons de la mort», «beaucoup plus ont été battus et torturés par des policiers en service» et «la violence continue», affirme Amnesty International (AI) dans un rapport publié jeudi à Londres. Selon l'organisation de défense des droits de l'homme, les enfants et adolescents contraints par la misère de leur famille à vivre dans la rue sont particulièrement vulnérables aux abus commis par la police

Gazette de Lausanne
September 5, 1990

FIGURE 1.3

Examples for agency mentions in newspapers.

coming from a news agency, on the other hand, can be located through the agency name. With the help of text similarity measures, this article can then be used to detect similar articles which probably stem from the same news agency release but which were not credited. Through this process, it is possible to extract many, although not all, of the newspaper articles with news agency content (Figure 1.2).

As a first step to finding all articles with agency content, this project focuses on detecting news agencies mentioned in articles in the corpus, as illustrated in Figure 1.3. The detection is done through a classifier (bidirectional encoder representations from transformers (BERT) model), which requires different steps from the exploration of the corpus, over the dataset construction towards the training of the model.

In the second part, the news agency detection model is executed on the whole *impresso* corpus, which makes it possible to conduct a first analysis of news agencies in the corpus and to address the research questions which motivated the initiation of this project. Thus, the project will treat the whole data pipeline, from the assembling of a data set over the training of a classifier to the inference of a big text corpus, which can be seen both as a strength and a challenge of the project.

With the analysis of news agencies in Swiss and Luxembourgish newspapers, this thesis contributes to the historical research on the relation between news agencies and newspapers and gives the opportunity to conduct a more detailed analysis of news agency content and all its different facets in the *impresso* corpus. On a technical level, the news agency classification will be incorporated in the metadata of the whole corpus, implementing a news agency filter as a search option on the *impresso* interface, hence opening the possibility to research news agency content specifically.

The main contributions of this thesis can be summarized in four points:

1. The building of a corpus of 1,530 articles annotated with news agency mentions;
2. The training and evaluation of several BERT-based models for the detection and classification of news agency mentions, and the release of two models (for French and German) identified as the most appropriate;
3. The analysis of the role of news agencies in the *impresso* corpus, based on the application of the two best agency detection models on the overall corpus;
4. The opening of new venues for research in text reuse.

The code and the dataset developed as part of the project are released under open licences and in line with open science best practices, with the usage of a collaborative software development platform with version control (GitHub)[5].

The remainder of the thesis is structured as follows. Chapter 2 introduces the background of the project, first giving an overview of the development of news agencies in the past, then presenting deep learning approaches for text classification and named entity recognition, before surveying studies on news agencies and their role in the media ecosystem. Chapter 3 details the different steps taken to construct the dataset used for the subsequent classification, which is the focus of Chapter 4. Next to the description and discussion of the conducted experiments, it motivates the choice for the French and German models, which were applied on the whole *impresso* corpus. A brief summary of the inference process followed by the analysis of its results can be found in Chapter 5. A general discussion of the project, its limitations, and an outline of future work and possible research directions in Chapter 6 concludes the thesis.

[5] `https://github.com/impresso/newsagency-classification`

2 Background

This chapter provides a general background to the master's project and an introduction to the relevant literature. On the historical side, we first give an overview of the development of news agencies in the 19th and 20th centuries to get an idea of the content of the study. Next, we focus on the technical side and introduce deep learning-based approaches for text classification and named entity recognition (NER). Finally, we present existing studies on the importance and role of news agencies in the media landscape.

2.1 News Agencies in the 19th and 20th Century

19th Century: Beginnings of News Agencies and the European Cartel. The first news agency, *Agence Havas*, was founded by Charles-Louis Havas in Paris in 1835. Coming from the business of translating newspapers for news sheets, Havas extended his services to include information from a network of correspondents all over Europe, which he had set up three years before (M. B. Palmer 2019, p. 9). By 1838, the agency delivered its summary of local and foreign news not only to newspapers from France, the Netherlands, Belgium, England and some German states but also to the French government (ibid.). Following the example of Havas, Bernhard Wolff and Paul Julius Reuter founded their own news agencies in Berlin (*Wolff'sches Telegraphenbureau*, 1849) and London (*Reuters*, 1851) respectively. Especially P. J. Reuter was interested in the technological side of the business, first specialising in transmitting financial news to the London market, before he broadened his offer to include general news as well (M. B. Palmer 2019, p. 15).

All news agencies profited from the technological advancements of the time. While the beginnings of their business adventures still included ponies, pigeons and railways, soon the electric telegraph became the preferred means of information transmission (ibid., p. 10). Nevertheless, telegrams were expensive, so despite their competition on the market, the agencies Havas, Wolff and Reuter decided to work together to save costs. In 1859, they sat together in Paris and agreed on spheres of influence for each agency, where the respective agency was obliged to collect information and distribute it to the other two. In exchange, it had exclusive rights to the market, thus protecting the agency's business. Thereafter, Havas' spheres of influence were France, Spain and Italy, South America as well as India and China. Wolff got Germany, Russia, Scandinavia and the Slavonic countries, whereas Reuters' influence extended over the British Empire and the Far East (Shrivastava 2007, p. 13). So, although a telegram was signed "Havas", it might as well have come from another agency such as Reuters (M. B. Palmer 2019, p. 12).

This European news agency cartel dominated the market and lasted until the 1930s, with a few changes over the years. In 1875, *Associated Press (AP)* joined the cartel. The American news agency had been founded in 1848 by newspapers from New York to save costs on the coverage of the US-Mexican war. As part of the cartel, it could tap into the news network in Europe, while concentrating on gradually consolidating its monopoly on the American market (M. B. Palmer 2019, p. 19ff).

Although the cartel dominated the trade with news, other agencies were founded in Europe in the 1850s and 1860s, including the Italian *Agenzia Stefani* (1853), the Spanish *Fabra Agency* (1867) and the British *Press*

Association (1868). Sometimes governments themselves would initiate the establishment of a national news agency, as was the case with *k. k. Telegraphen Korrespondenz-Bureau* (1860) in Austria-Hungary and *Balgarska telegrafna agencija* (BTA, 1898) in Bulgaria. Outside the cartel, news agencies were mainly operating nationally and thus, even though not part of the big three, had contracts with Havas, Reuters or Wolff to supplement their national service with foreign news (Boyd-Barrett and M. Palmer 1981, p. 426). In some countries, the three agencies also opened their own agency branches (e.g. in Luxembourg), bought up rivals or bound national agencies to them. The latter was for example the case in Italy, where Havas had exclusive access to news from Stefani, even letting Stefani send Italian dispatches in the French agency's name to other countries like Germany, Switzerland or Great Britain (ibid.). Additionally, national agencies had agreements with their counterparts in other countries (Boyd-Barrett and M. Palmer 1981, p. 436). Thus, the European agency landscape can rather be seen as a big network of news actors than a fragmented market.

In Switzerland, although the country was not part of the cartel's agreement, the major players were Havas and Wolff – but only until 1894, when editors of several newspapers founded the cooperative *Agence Télégraphique Suisse (ATS)*, in German *Schweizerische Depeschenagentur (SDA)*. Their goal was to have an independent source of information and not to rely on the two foreign news agencies. By 1900, ATS already had 71 subscribers, which amounts to almost the entire press in Switzerland (Shrivastava 2007, p. 5f). In 1916 and 1917, two other (multilingual) Swiss news agencies joined the market, namely the *Schweizerische Press-Telegraph (SPT)* and the *Schweizer Mittelpresse* (SMP, from 1944 until 1993 named *Schweizerische Politische Korrespondenz SPK*[1]) (Meier 2010).

To sum up, the 19th century featured the foundation and rapid growth of news agencies in Europe and the US, where the agencies became key actors in the news world. In one of the first studies on the news market in Germany, Groth (1928, p. 22) (cited in Terhi Rantanen 2019) defined their role as follows:

> *"News or telegraph agencies are correspondence bureaus, whose special characteristic lies in the transmission of news reports. They are enterprises which systematically gather news in the fastest possible way and, after reviewing and editing this, transmit it to newspapers and other interested parties in the most rapid manner possible."*

20th Century: The End of the Cartel, World War II and a Changing World. At the beginning of the 20th century, a new player appeared in the American market: *United Press (UP)*, privately founded in 1907, aimed to break the supremacy of *AP*. Together with the rise of the United States on the world stage, UP's influence grew, as it endeavoured to use the developing American cable network around the world to set up its own network of correspondents (Boyd-Barrett and M. Palmer 1981, p. 438). UP rejected the offer of Reuters to replace AP in their agreement and found allies with the small news agencies which were unhappy with the oppressive power of the cartel. Under pressure by UP, AP broke the contract of the cartel by concluding contracts with national agencies like *TASS* (Telegrafnoe Agentsvo Sovetskogo Soiuza, Telegraph Agency of the Soviet Union, founded in 1925), which lay in the influence area of its allied agencies. Although the bounds between the big agencies had already been weakened before, the cartel dissolved officially in 1934, when AP formally left the cartel (Shrivastava 2007, p. 15).

Whilst already having a more or less close relationship with their respective governments during peaceful times, news agencies, especially the major ones, often became mouthpieces of their country during the war (Shrivastava 2007, p. 7f). Under the Nazi regime, Wolff was put under total control, even being merged with the German *Telegraphen-Union* in 1934 to form the news agency *Deutsches Nachrichtenbüro (DNB)*. Havas suffered the same fate during the German occupation of France, when its agency was changed to *Office Français d'information (OFI)*. Switzerland, on the other hand, became an international news hub with many newly founded agencies. However, none of them survived for long (Meier 2010).

[1] In French: Correspondance Politique Suisse (CPS)

With the reorganization of Europe after the war, the landscape of news agencies also adjusted. Next to the French *Agence France Presse (AFP)*, which was already instated in 1944, 24 new agencies set up their operations between 1945 and 1949 (Shrivastava 2007, p. 18). Among them were the German *Deutsche Presse-Agentur (DPA)*, the Italian *Agenzia Nazionale Stampa Associata (ANSA)* and the *Austria Presse Agentur (APA)*. The same phenomenon could be observed during the wave of decolonization in the 50s and 60s in Africa, when most of the freshly established states also started their own national news agency. The Arabic states joined this development: a news agency, as Terhi Rantanen (2019, p. 3) puts it, "became a national symbol like a national library, bank, or post office, serving the media in their national language(s)".

On the international stage, the news agencies Reuter, AP, UP and AFP continued to play major roles. Yet, in 1958, UP merged with the *International News Service* and became *United Press International (UPI)*, which preceded its slow decline from power (M. B. Palmer 2019, p. 141ff). In Switzerland, ATS and SPK (until 1993) dominated the market, although UPI (1961-1972), the *Deutscher Depeschen-Dienst (DDP)* (1972-1983) and AP (since 1981) were also present (Meier 2010). The last significant change in the European agency landscape happened after the fall of the communist regimes in Eastern Europe, which resulted in the formation of new, private agencies and the reorganization of existing ones (Shrivastava 2007, p. 30).

Apart from the political changes, technological advancement challenged (and continues to challenge) the role of news agencies in the media environment. While agencies managed to adapt to inventions such as the telephone, and the transmittal of news in the form of images or videos, the internet fundamentally changes the rules of the game: Now, in the 21st century, news is fast all over the world, everyone can get international information without having to rely on news agencies or their main clients, the traditional media. This causes the need for the whole news sector, and especially news agencies, to reinvent themselves (Terhi Rantanen 2019).

2.2 Deep Learning Approaches for Text Classification and Named Entity Recognition

This section gives a general overview of the development of deep learning (DL) in natural language processing (NLP)[2], before concentrating on the state of the art of the two tasks which are important for these projects, namely text classification and named entity recognition.

Deep Learning in Natural Language Processing When machine learning (ML) models entered the field of NLP, it was part of a two-step process. At first, experts with domain knowledge hand-crafted features from the given data, which secondly were fed to an ML model such as the hidden Markov models (Rabiner and Juang 1986) or support vector machines (Cortes and Vapnik 1995) for classification. In 2013, Mikolov, K. Chen et al. (2013) proposed the word2vec algorithm, which allowed to get a feature per word (known as *word embeddings*) not through expert knowledge, but on the basis of the usage of the word in a relevant text corpus (Mikolov, K. Chen et al. 2013). More explicitly, a simple feed-forward neural network was trained to predict a word based on its surrounding word window, and the hidden states of the network serve as the word embeddings. As those embeddings managed to capture the meaning of words in a vector space (Mikolov, Yih and Zweig 2013) and are easily used as input for ML algorithms, they were soon widely used and improved further (Pennington, Socher and Manning 2014).

Progress in the field also benefited from the proposal of deep learning (DL) architectures such as recurrent neural networks (RNNs), which allowed to train models end-to-end, i.e. with a text sequence as input and

[2]This overview stays on a coarse-grained level; for more detailed explications and further references see Lauriola, Lavelli and Aiolli 2022.

a classification or text generation as output. Although RNNs were devised to remember the words they had seen before through their hidden state, they had problems with long-term memory, so models based on the architecture of RNNs, but with explicit memory management such as long short-term memory (LSTM) (Hochreiter and Schmidhuber 1997) and gated recurrent unit (Cho et al. 2014) were introduced. Together with convolutional neural networks (CNN, Collobert et al. 2011), which were originally developed for image processing, they became the state-of-the-art of NLP (Yin et al. 2017) and are still used now (e.g. Wang et al. 2021).

In 2017, Vaswani et al. (2017) stirred up the field of NLP with their transformer architecture. The model relies heavily on the attention mechanism, an algorithm which models dependencies of a word on other words in its context, independent of their distance to the word in question. Because of its non-sequential nature, the transformer is easily parallelisable and thus scalable. The original model had an encoder-decoder structure, as it was originally proposed for the task of machine translation, but models using the encoder and decoder separately also had remarkable success, such as bidirectional encoder representations from transformers (BERT) on the encoder side (Devlin et al. 2019) or generative pre-trained transformers (GPT) and its successors on the decoder side (Radford and Narasimhan 2018). Those models now widely serve as pre-trained foundation models (Zhou et al. 2023), meaning that they were first trained on large unlabelled text corpora through the prediction of a most likely next word in a sequence or a context, and can subsequently be fine-tuned on a smaller, annotated corpus for the target task (also called transfer learning, see Ruder et al. 2019).

Text Classification. Minaee et al. (2021, p. 62:1) define the task of text classification as follows: "Text classification, also known as text categorization, is a classical problem in natural language processing, which aims to assign labels or tags to textual units such as sentences, queries, paragraphs, and documents." In their paper, they provide a comprehensive survey of the deep learning models used for text classification, showing that still a great variety of model architectures is used, ranging from LSTMs, CNNs or graph neural networks over memory-augmented networks or hybrid models towards transformers (ibid.).

Recently, one of the most influential models is BERT (Devlin et al. 2019). The encoder structure of the transformer architecture takes the whole text sequence as input at once (as opposed to looking only at a word and its past), which makes it predestined for tasks like text classification. BERT's architecture is based on a stack of 12 or 24 identical blocks, depending on the size of the final model. Each block consists of a multi-head attention module, which either takes the (sub-)word embeddings and their positional encoding from the input or the output from a previous block and employs the attention mechanism on it multiple times in parallel (thus *multi-head*). After the addition of a residual connection (K. He et al. 2016) and a normalization of the output (Ba, Kiros and Hinton 2016), the resulting vectors are given to a feed-forward network. Another output normalization and residual connection then complete each block. The parameter count of the two original BERT models amounts to 110M or 340M parameters, with the bigger version scaling up the number of blocks, the hidden vector size and the number of self-attention heads.

Since its proposal in 2019, many variants of BERT were published, aiming to make BERT either more performant or more efficient. Among them are RoBERTa (Liu et al. 2019), which uses more training data and prolongs pre-training, ALBERT (Lan et al. 2020), which deploys parameter reduction techniques, resulting in lower memory consumption and increased training speed, or DistillBERT (Sanh et al. 2020). The latter introduces knowledge distillation techniques to reduce the size of BERT by 40%, hence making it faster, but still retaining 97% of its language understanding capabilities. Another approach is to extend BERT by including knowledge from external knowledge bases (ERNIE, Sun et al. 2019).

Named Entity Recognition. The task of named entity recognition (NER) comprises the identification of named entities such as locations, organisations or persons in a given text. For this, models classify every

word (token) in a text, either with the respective named entity category or with an *O*, if it does not constitute a named entity. Because the classification of words is at the heart of NER, good word embeddings are a crucial part of the success of a model (see e.g. Wang et al. 2021). Instead of or additional to the word embeddings already introduced above, researchers also use character-level embeddings, usually computed through a CNN (Dos Santos and Zadrozny 2014), or sub-word embeddings like *fastText* (Bojanowski et al. 2017). They have the advantage of being able to deal with previously unknown or misspelt words. One problem of NER is word disambiguation, where a word can have the same spelling, but completely different meanings, which can only be determined through its context. To tackle this, Akbik, Blythe and Vollgraf (2018) propose character-level contextual string embeddings, where a word can have different embeddings based on the context it appears in. Peters et al. (2018) follow the same principle with their embeddings from language models (ELMo), but they use the internal states of an LSTM as embeddings. The idea of taking internal states of a language model as embeddings has since also been applied to transformers (Schweter and Akbik 2021).

Concerning model architectures for NER, LSTMs have long been at the forefront (Lample et al. 2016, Yadav and Bethard 2019), but are now slowly being overtaken by the transformer architecture (Boros et al. 2020). Commonly, another layer is put on top of the model, called conditional random fields (CRF) (Lafferty, McCallum and Pereira 2001). CRF is a sequence modelling framework which makes predictions of an entire sequence instead of only one word, hence taking into account dependencies between labels in a sequence – an important aspect during NER. However, in their study comparing an LSTM-CRF, a fine-tuned transformer and a transformer-CRF, Schweter and Akbik (2021) found that the transformer performs better than the LSTM and that the CRF layer on top of the transformer does not make a big difference for the result.

The subfield of NER on historical data roughly follows the trends of NER in general, i.e. transformer architecture slowly replaces LSTM-CRFs and various forms of embeddings (character, sub-word and word granularity, contextualized or in-domain embeddings) are of great importance (Ehrmann, Hamdi et al. 2023). However, historical NER is faced with many difficulties specific to its domain. In their survey on the state of historical NER, Ehrmann, Hamdi et al. (2023) identify four key challenges: (1) the need to deal with a great variety of data, such as different document types, time periods, topics; (2) noisy input, stemming from faulty OCR or OLR; (3) dynamics of language, including historical spelling variations, naming conventions or entity and context drifts; and the (4) lack of resources to train NER models. According to the authors, possible solutions include the use of transfer learning, OCR post-correction, training of historical language models and the building of more in-domain training data.

2.3 Existing Studies on News Agencies in the Media Ecosystem

News agencies and their role in the media ecosystem have been the focus of researchers for some time (Terhi Rantanen 2019), the first study going back as far as 1928 (Groth 1928). They often concentrate on a single agency, relationships among a group of agencies, or the news situation in one country (J. He 1996; T. Rantanen 1990; Schwarzlose 1989; Silberstein-Loeb 2014). Some studies also examine the news agencies' role in news production (Boyd-Barrett 2000; Czarniawska-Joerges 2011) or news flows, e.g. how press releases are ingested by news agencies and passed to the newspapers (Forde and Johnston 2013). First quantitative studies on the amount of news agency content in newspapers have already been conducted in the 1950s (Institut international de la presse 1953; Schramm 1959), when institutions such as UNESCO took an interest in news agencies and their role in the changing news world (UNESCO 1953). In their study on news circulation in the US, India and eight European countries[3], the Institut international de la presse (1953) inspected 177 newspapers and 45 daily records from news agencies. In an additional

[3]Belgium, France, Germany, Great Britain, Italy, Netherlands, Sweden and Switzerland

survey with the directors of several European newspapers, they found that on average, 40-70% of the foreign news published in the newspapers came from news agencies, while approximately one-fifth to one-third of editorial space was dedicated to foreign news.

Another influential study on the origin of news in the media was conducted by Lewis, Williams and Franklin (2008), who manually evaluated 2609 items from five British national "quality" newspapers as well as radio and television news reports in 2006. Their results led them to the conclusion that "meaningful independent journalistic activity by the media is the exception rather than the rule" (ibid. p. 17), inter alia finding that around half of the articles were based on agency content, with 30% of all articles being verbatim copies, but only 1% were attributed to a news agency.

With the emergence of digital sources and tools, researchers could explore new methodologies when analysing news agency content in the media. Vogler, Udris and Eisenegger (2020), for example, attempted to measure media content concentration based on text similarity in seven newspapers between 2012 and 2018 in German-speaking Switzerland. A subsequent manual content analysis showed that 33.5% of all articles (n=13,993) contained a reference to a news agency. With Welbers et al. (2018) and Boumans (2018), two research teams published large-scale quantitative analyses on the influence of agencies on the Dutch news market. Welbers et al. concentrated on the influence of the Dutch agency ANP on the leading newspapers and online media in the years 1996, 2008 and 2013. Using a measure of document similarity between (political) articles and the releases from ANP, they could trace around 31% back to the agency, and around 9% of the articles were verbatim copies. With a similar methodology, but extended by a more detailed examination of text reuse, Boumans (2018) showed that the ratio of agency content is significantly higher in online media (up to 75%) than in print media (between 12% and 48%). Apart from one newspaper which did not credit agencies at all, the ratio with which an agency was acknowledged as the source was rather high in the printed newspapers (70% and 94%). In another large-scale study, Nicholls (2019) looked at the media in the US and Great Britain, detecting shared text between different news articles in the time from April 13 to May 12 in 2017. For the collection of the corpus, they used databases and web scraping, assembling 163,297 news articles, 125,508 articles from four major news agencies and 24,675 press releases from both countries. In contrast to the findings from Lewis, Williams and Franklin (2008), they only detected a textual overlap with an agency release in 27.3% of the articles.

Although recent research has looked at the influence of news agencies on the content produced in newspapers (and online media) on large corpora of text, to our knowledge, the present project is the first large-scale quantitative study of news agency content in the 19th and 20th centuries. However, unlike the previously presented research, we do not have access to agency releases and therefore have to rely on the attribution of agencies in the articles and their presence in text reuse clusters.

3 Data

This chapter details the steps and reflections which led to the dataset used for the training and evaluation of the classifier. For this, relevant data needed to be collected in *impresso*, followed by sampling, annotation and conversion into the final data format. An overview of the different datasets existing throughout the process is displayed in Figure 3.1.

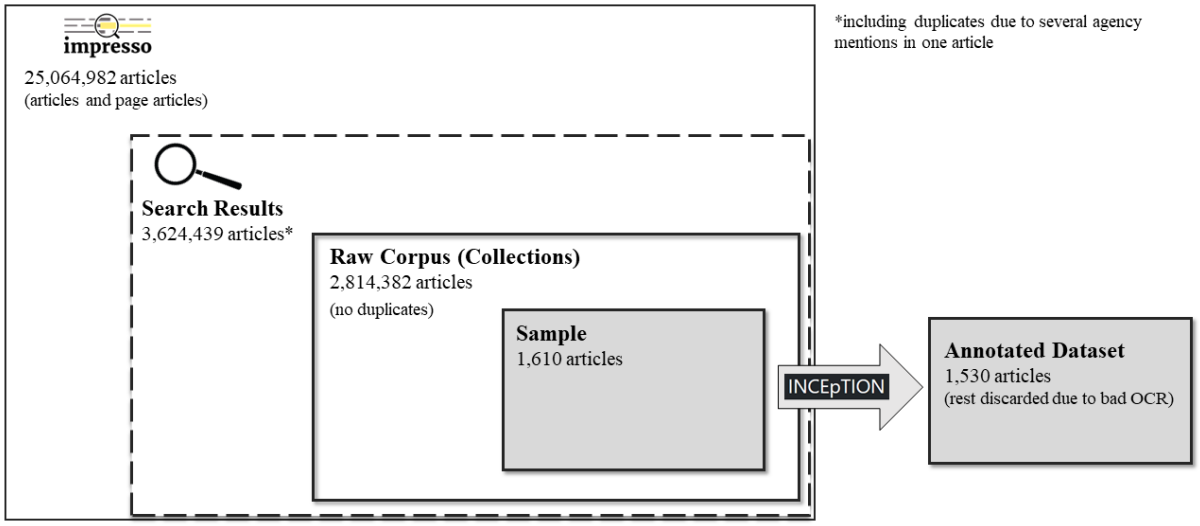


FIGURE 3.1
Overview of the different stages of the dataset's construction.

To get from the search results in *impresso* to the final dataset for classification, several platforms and APIs were used, which each worked with a different data format, requiring us to transform the data for each consecutive processing step.

This data wrangling pipeline was already set into place in 2020 for the HIPE annotation campaign (Ehrmann, Romanello, Flückiger et al. 2020), which facilitated the format conversions enormously and reduced the workload for this task. The data conversion therefore mainly involved understanding the existing code and modifying it to suit the particularities of this project, allowing the focus to be on the construction of the dataset.

The following sections describe the steps and considerations throughout the building of the dataset, starting with a definition of the elements to be considered for the dataset in Section3.1, then moving on to the sampling process and annotation in Sections 3.3 and 3.4, before providing an overview of the final dataset in Section 3.5.

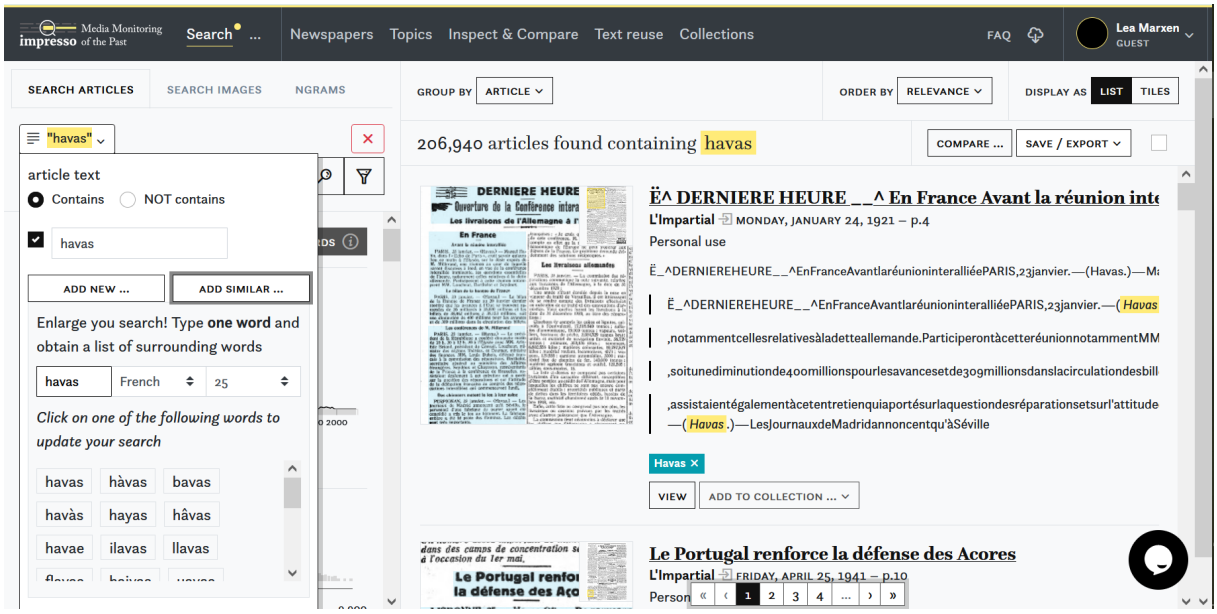


FIGURE 3.2
The *impresso* interface, showing the results for a simple "Havas" query.

3.1 Working Definition for a News Agency Article

As already outlined in the introduction, the *impresso* corpus does solely contain newspaper articles and not agency releases, which obliged us to detect agency content with the help of references to agencies in the articles, with subsequent enrichment through text reuse clusters.

For the building of the training corpus, we thus chose to concentrate on articles we could confidently identify as having agency content, which motivated the following working definition:

> *An article with news agency content is anything that explicitly cites a news agency as a source.*

This offered us a good starting point for the collection of relevant data in *impresso*.

3.2 Building a Raw Corpus in *impresso*

First, we collected articles which might mention a news agency as a source (subsequently also called "news agency articles"). This was done by querying the name of news agencies in the *impresso* interface, which can be seen in Figure 3.2.

impresso provides the functionality to save all articles found through a query as so-called "Collection". This made it possible to create one collection for every queried news agency. In some cases, if a collection was too large, it had to be split up into several ones. The resulting compilation of the collections served as a raw corpus from which the training data was sampled in the next step.

3.2.1 Querying for News Agencies in *impresso*

In order to build the news agency collections in *impresso*, two questions had to be answered: (1) Which agencies should be queried? And (2) what should the queries look like?

Choosing Relevant News Agencies

To answer the first question, we originally endeavoured to list all news agencies which have ever existed and to query all of them. Although there are only a handful of news agencies which operate internationally, the number of agencies on a national and regional level made it quickly clear that it was neither feasible nor expedient to consider all of them (an unfinished list contained roughly 200 agencies). Thus, we focused on those agencies which might appear in the *impresso* corpus, namely European and internationally operating news agencies. To find all relevant agencies, we conducted literature research, drawing from historical research (Boyd-Barrett and M. Palmer 1981; M. B. Palmer 2019; Terhi Rantanen 2019), encyclopedia (F. A. Brockhaus 1894; Meyer 1909) and Wikipedia (Wikimedia Foundation 2023). We complemented this by an exploratory search through the *impresso* corpus, which encompassed samples of both Swiss and Luxembourgish newspapers over all decades in the 19th and 20th centuries.

The literature and exploratory search resulted in around 70 news agencies which were queried in *impresso*. The list can be found in the Appendix, Table B.1.

Query Design

The text that can be queried in the *impresso* interface contains errors which stem from a variety of challenges during the digitization process of the newspapers, such as the age of the newspapers, the quality of the images or the performance of the OLR and OCR algorithms. To account for this, it is possible to enrich a query with similar words, i.e. words whose *impresso* word embeddings are close to the embedding of the queried word.

We heavily used this functionality for the news agency queries, alongside synonyms or abbreviations of the respective agencies. For example, the query for the agency *Wolff'sche Telegraphenbureau* looks as follows:

```
containing wolff or lwolff or lwolsf or lwolss or iwolff or wofff or
wolffsbüro or wolffbüro or wolffbüros or wolffbureau or wolffbureaus
                    or wolff'sche or wolffmeldung
```

Often, brackets around the agency mentions are mistaken by the OCR for letters, which explains search-words like *lwolff* or *wolfff*.

Additionally, the queries are automatically processed by the *impresso* search engine. To cite a contributor of *impresso*: *"At indexing time, the search engine does a standard tokenisation, which also implies: lower casing + the application of a basic stopword filter +, for German, the application of a light stemmer and token 'normalisation' for handling multiterm tokens (compounds). The query which the user enters undergoes the same process."* This means that in German, querying for e.g. *Wolffsbüro* includes matches for *wolffsburo* as well.

As a last step, we only saved news agency queries yielding above 2,000 search results in a collection. If there was a high risk of too many false positives, i.e. articles which responded to the query but did not contain a news agency, we also abandoned query results with slightly more than 2000 articles. The threshold of 2000 was chosen as a trade-off between having enough news agencies to make the corpus representative, while at the same time ensuring a compact corpus which is easy to work with subsequently.

The different collections and their queries can be found in Appendix B, Table B.2.

3.2.2 Raw Corpus Statistics

The Collections were downloaded as CSV files, with one row per article. The corpus contains 27 collections, one for each agency, amounting to 2,814,382 articles in total. Figure 3.3 shows the number of

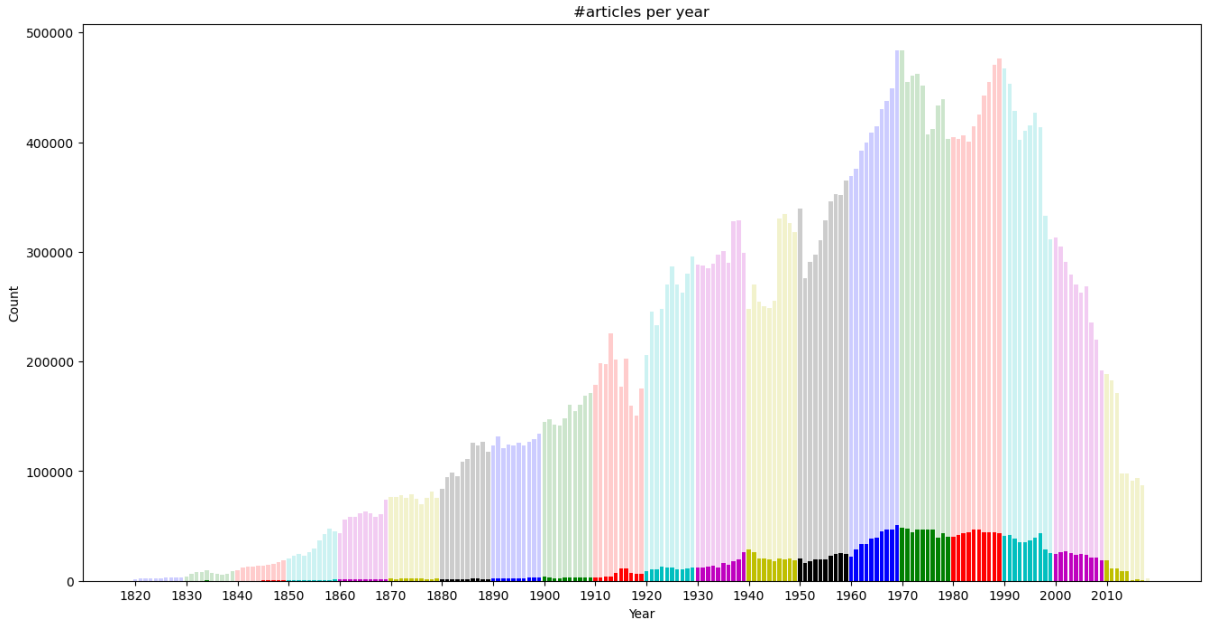


FIGURE 3.3
Number of articles in the raw corpus per year, compared to all articles per year in the *impresso* corpus (light colour).

articles in the raw corpus (Collections) against the number of all articles in the *impresso* corpus per year. Before 1910, the yearly ratio of articles in the raw corpus vs. the articles in *impresso* does not exceed 3%, but afterwards, it slowly increases and reaches its highest number in the 1960s to 1980s with around 10%. Although these numbers are only rough and include many false positives, they already point to a tendency of increased usage or attribution of content to news agencies in the latter decades of the corpus.

Figure 3.4 displays the size of the different Collections, showing that only a few agencies often appear in the corpus. However, the presence of false positives might skew the distribution towards agencies whose acronyms are very generic. This could for example explain the high number of articles with *AP*, as the two-letter acronym might be easily used for many abbreviations, and is prone to appear during wrong word splits due to OCR errors (e.g. splitting *appeler* into *ap* and *peler*).

In general, 94% of the articles in the raw corpus comes from Swiss newspapers, whereas the Luxembourgish newspapers only make up 6%. The languages are also unevenly distributed, with 78.07% of the articles being French, 21.88% German and 0.05% Luxembourgish. A handful of articles are also declared to be English, although this is very probably due to misclassification, as the *impresso* corpus does not (yet) contain English articles.

3.3 Sampling

The raw corpus was too big to annotate, so a representative corpus needed to be sampled. The following section details the sampling process, from the choice of the sampling strategy to the application of thresholds towards the actual sampling.

3.3.1 General Sampling Strategy

The goal of the dataset construction was to have a training corpus which optimally trains the classifier, such that it generalizes well when it is applied to the whole *impresso* corpus. For this, the data needed to

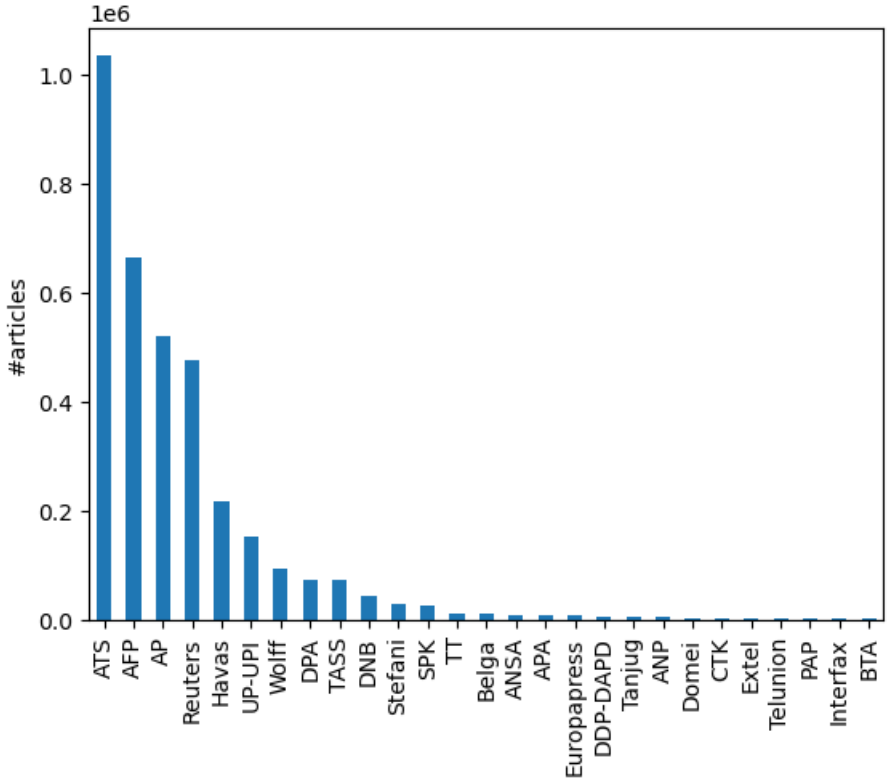


FIGURE 3.4
Number of articles in each agency collection from the raw corpus.

mirror the variety existing in the whole corpus, but at the same time, every element should be represented sufficiently to give the classifier the possibility to learn from it at all.

To accomplish this tradeoff, we applied a mixture of uniform and stratified sampling, namely:

- uniform over decades: 100 articles per decade;
- stratified for agencies & newspapers.

If an article contained several agency mentions, one of the mentions was chosen at random. Then, for each decade, the sampling of articles was conducted in proportion to the ratio of agency mentions and newspapers of the respective decade.

For the split of the articles in train/dev/test datasets, we chose the proportion 80%/10%/10%. We used stratified splitting for decades and agencies but did not take the distribution of newspapers into account, due to the comparatively small amount of articles in the dev and test sets.

3.3.2 Thresholding

Before we performed the sampling, we applied several thresholds to ensure meaningful and concise data.

Selection of Years

In terms of years, only articles which appeared between 1840 and 2000 were considered for the sampling. The lower bound is on account of the first news agency (Havas) being founded in 1835, while we chose

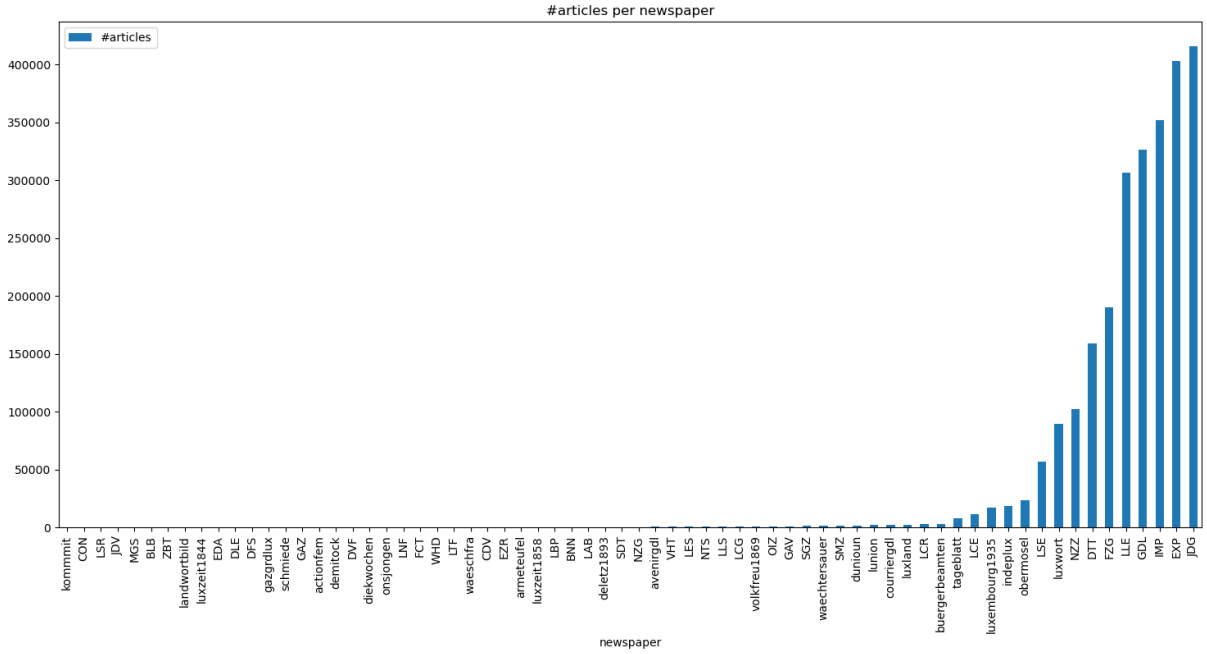


FIGURE 3.5
Number of articles per newspaper in the raw corpus.

the upper bound because this project was mainly interested in the development of news agencies in the 19th and 20th centuries.

Selection of Newspapers

In order to guarantee a smooth stratified sampling over newspapers, we applied a threshold beforehand to discard the newspapers with a very small number of articles in the raw corpus. Figure 3.5 gives an overview of the number of articles from every newspaper in the raw corpus.

As a prior step, the newspaper *Neue Züricher Zeitung* was excluded, as the textual content was not provided in the desired units of articles, but newspaper pages. An analysis of the remaining newspapers showed that out of the other 67 newspapers in the raw corpus, 10 newspapers featured less than 50 articles, seven newspapers had 50–75 articles and another three newspapers had 75–100 articles in the raw corpus. Additionally, because of the stratified sampling, only newspapers which made up at least 1% of all articles of a decade could be considered at all. As Figure 3.6 shows, selecting either all, the largest 50 or 40 newspapers only changed the choice of newspapers during the sampling of the first two decades.

Based on these results, we kept newspapers which had at least 75 articles in the raw corpus, which amounted to 50 newspapers in total.

Thresholding on Article Length

Regarding the article length, we applied a lower threshold to achieve a certain comprehensibility of the text. In the raw corpus, 0.5% of the articles were shorter than 32 tokens, whereas 1% were shorter than 39 tokens. Comparing those two thresholds by printing examples, see Figure 3.7, we can observe that articles shorter than 32 tokens seldom contain news agency content, and if they do, the article is cut off wrongly. For articles with length between 32 and 38 tokens, some articles still do not contain relevant content, but a significant number of articles does. This motivated the choice of the lower threshold of < 32 tokens.

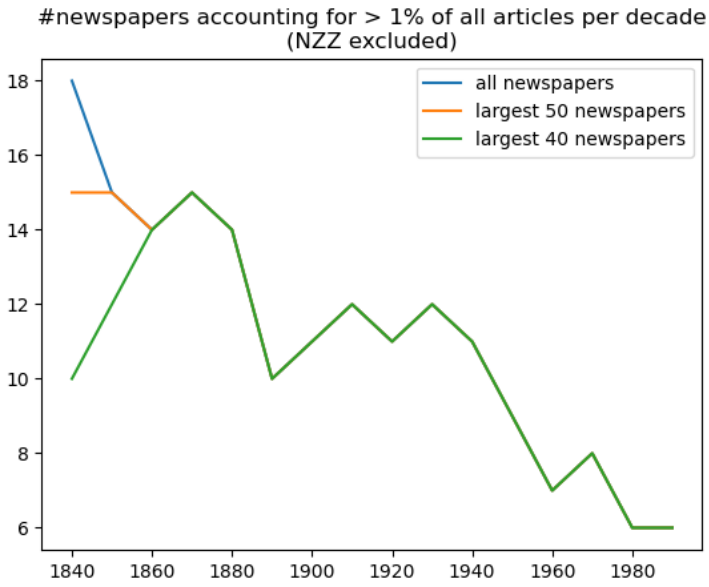


FIGURE 3.6

Due to the stratified sampling, only newspapers which feature at least 1% of all articles within a given decade are considered during the sampling process. This figure displays how many different newspapers would be chosen during the sampling process if the biggest 40, 50 or all newspapers were considered during sampling.

Articles with length < 32 tokens

- âp italanlagyn zu 4 , 412 bis 5 Prozent vermittelt das BanlComptoir von in Seanfs .
- DDP— AbonneMents HW aui oie „ Vunoner- « acyricycen nno oeren „ Wochenblatt werden fortwährend angenommen . Erschienene Nummern des „ Alphorn werden gratis nachgeliefert .
- M. Torrcnté, de Massongcx, offre à louer un ap partement et un four tout neuf. S'adresser à lui même pour le prix et les conditions.
- A LOUER, pour la Si-Martin prochaine, un ap partement, rue de Conthey, à Sion. S'adresser au gérant du journal. SION. — IMPRIMERIE DE DAVID RACHOR
- **Konstantinopel. Nach den Informationen des Wolffbureaus hat die Pforte beschlossen, die russische Forderung bezüglich der Oeffnung der Dardanellen zurückzuweisen, da fönst land eine urädominierende Stellung in Konstan**

Articles with length between 32 and 38 tokens

- QemütlicKer Vereinigung 3 am 8 taß, clen 7. im Dezember l 912 Notel „ LaKnliof" in Ker êrg sreun 6 ! icb 8 t laclet ein ttermnnn lirnmer, Wirt ŝsZI'KMK Î'mit
- **USA Die Rede des Präsidenten Roosevelt Neuyork , 11 . . luni . (Reuter .) Präsident Roosevelt hielt arn Montagabend in Charlotteville eine Rede . Er führte u . a . aus , die amerikanische Regie-**
- **Das Tschechische Nationalkomitee In London London , 16 . Juli . (United Preß .) Man erwartet hier , daß England noch diese Woche das Tschechische Nationalkomitee in London als legale Regierung der Tschechoslowakei anerkennen werde .**
- Karfreitag i Ungar . Champignons , Südungarische Eier Karsamstagt Ungarisches Paprika-Huhn Ostersonntag t Slebenbürger Gulasch , Jungfrau-Braten Ostermontag t Tschlkosch Rostbraten , ung . Paprika-Huhn TägUthi Gulasch en tasse (OC) , Palasclnta
- **Kairo , 22 . April . (United Preß .) Das britische Hauptquartier gibt in einem Communlquö bekannt , daß die britischen Truppen in Griechenland nunmehr Im Süden von Lamia Verteidigungsstellungen bezogen hätten .**

FIGURE 3.7

Examples for articles with length of less than 32 tokens and between 32 and 39 tokens respectively. It can be observed that the latter articles exhibit more comprehensible sentences with more news agency references.

We set the upper threshold to 2,000 tokens, thereby excluding around 4% of the articles (not including the "page articles" of the NZZ). A look into some articles with 2,000 tokens shows that they occupy more than half a page, sometimes even a whole page, which suggests that the OLR did not correctly determine the article boundaries in most of these cases. Thus, this threshold is not too conservative, keeping some "dirty" data the classification model will be confronted with.

All in all, because of the thresholding, 8% of the articles were dropped, leaving 2,308,268 articles in the corpus to sample from.

3.3.3 Sampling Process

The sampling was conducted on the thresholded raw corpus according to the sampling strategy proposed in Section 3.3.1. In order to ensure that a news agency can be present in the training, dev and test datasets respectively, we imposed an additional constraint: For each decade, a news agency should appear in the sample at least three times. If this was not the case, additional articles were sampled. However, one has to keep in mind that the articles designated to contain a certain agency mention could be false positives and thus might not actually feature the anticipated agency mentions. This would only become evident after the completion of the annotation.

In the end, the sample contained 10 supplementary articles because of the added constraint. Thus, a total of 1610 articles were sampled (100 articles per 16 decades plus the 10 additional ones), with a train/dev/test split of 1289/160/161 articles respectively.

The sample contains articles from the following news agency collections:

`AFP, ANSA, AP, ATS, Belga, DPA, DDP-DAPD, DNB, Europapress, Extel, Havas, Reuters, SPK, Stefani, TASS, Telunion, UP-UPI, Wolff`

Some agencies were not sampled at all, because they did not appear often enough in any decade. They are:

`ANP, APA, BTA, CTK, Domei, Interfax, PAP, Tanjug, TT`

Moreover, the sample includes 37 newspapers. 1,203 articles are in French, 406 in German and one is classified as Luxembourgish, although after inspection turned out to be in French, too. In terms of country, 1,233 articles are from Swiss newspapers, while 377 articles come from the Luxembourgish press. A comparison of the sampled data to the raw corpus can be found in Figure 3.1.

TABLE 3.1

Table with statistics on the raw and sampled dataset, respectively. The column of possible agency mentions indicates the number of search hits for the agency queries and thus includes false positives.

Dataset		Agencies	Articles	Possible Agency Mentions	Tokens	Timespan
Raw	de	27	615,789	798,225	778,725,897	1780-2010
	fr	27	2,197,041	2,704,550	978,826,852	1738-2018
	en & lux	19	1,552	1,587	501,161	1842-2007
Total		27	2,814,382	3,504,362	1,758,053,910	1738-2018
Sampled	de	17	406	481	251,149	1840-1999
	fr	19	1,203	1,387	629,214	1840-1999
	lux	1	1	1	154	1947
Total		22	1,610	1,869	880,517	1840-1999

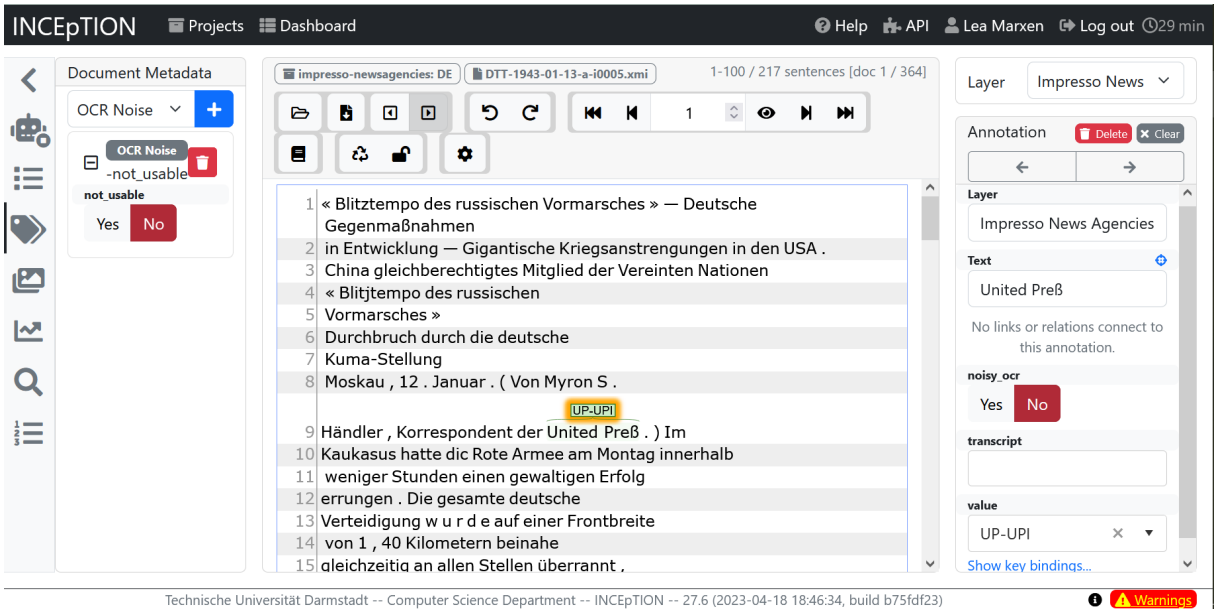


FIGURE 3.8
The Inception Interface during annotation: The text to annotate can be seen in the middle, while metadata annotation is done on the left and token-level annotation is specified on the right.

3.4 Annotation

As a final step in the dataset construction, the previously sampled 1610 articles were annotated to provide the classifier with a ground truth it can learn from. This section introduces the annotation tool *Inception* before presenting the annotation pipeline including the annotation setup and process, the inter-annotator agreement and the post-processing conducted afterwards.

3.4.1 Inception

For the annotation, we chose the web-based annotation platform Inception[1] (Klie et al. 2018). A screenshot of the annotation interface can be seen in Figure 3.8.

Inception is designed to support a wide range of annotations, including more demanding tasks like concept linking or knowledge base population (Klie et al. 2018). Although this project only needed token-level and document-level annotations with pre-defined tag sets, the platform was suitable because of several reasons. Firstly, as the DHLAB had already been working with Inception during their HIPE annotation campaign (Ehrmann, Romanello, Flückiger et al. 2020) in 2020, this project could rely on existing project templates in Inception as well as code to prepare the data for Inception, to download the annotated articles and to post-process them. While the templates and code needed to be adapted, they provided a very useful base to start from. Secondly, next to basic functionality like annotation, monitoring, curation (to resolve differences in annotation between several annotators) or evaluation tools, Inception offers a set of tools which facilitate annotation. They include the functionality to report articles if an annotator encounters problems during the annotation, the ability to show the image segments next to the OCR transcriptions as well as recommendation algorithms, which propose annotations on the basis of the prior activity of an annotator.

[1] Inception version 27.6

3.4.2 Annotation Setup and Process

The annotation setup in Inception included four separate projects, which only differed in the data they contained: For each language (i.e. French and German), we set up two projects. The smaller of the two projects included 48 articles, three for each decade, and served the purpose of computing the inter-annotator agreement (IA agreement). The other project contained the rest of the articles in the respective language ("core corpus").

Possible Annotations

The main annotation consisted of the token-level annotation of news agency mentions. For this, we created a tag set of possible news agencies, which comprised 27 news agencies (see Table C.1 in the Appendix). Additionally, we supplied an *unk* tag to deal with news agencies not contained in the tag set, as well as the tag *pers.ind.articleauthor*, which is discussed below. Shortly after the start of the annotation campaign, we added the tag *ag* for the generic *ag./Agence/Agentur*, as it appeared frequently in the articles.

After having tagged a token as a news agency mention, it could also be flagged to have a noisy OCR, with the possibility to write the corrected version of the mention in a designated transcript field (see Figure 3.8, right column). If the whole article was too noisy to be properly understood, it could be tagged as `not_usable` on the document level (see Figure 3.8, to the left), to be discarded afterwards and thus not used for the following processing steps. This option became necessary because the articles were not manually triaged beforehand.

The document-level annotation, stating if an article has news agency content or not, was excluded from the annotation in Inception, as it could later be inferred from the annotated agency mentions.

Annotation Guidelines

In order to ensure homogeneous annotations among the different annotators, annotation guidelines were provided (see Appendix C.2). In addition to an overview of the annotation task, the possible annotations and the handling of OCR noise, they contained clarifications on some anticipated difficulties. The first difficulty can be seen in Figure 3.9 (1): An author of an article (right column) might be credited the same way as a news agency (left column). To account for this, we introduced the tag *pers.ind.articleauthor*. If a case was unclear to an annotator, they could tag it as *unk* and report the article, which we then double-checked during the post-processing.

For the treatment of compounds and proper names, the annotation guidelines of the HIPE 2020 project were applied. For example, in Figure 3.9 (2), only the *Reuter* in *Reutermeldung* would be annotated. Finally, the use of the period needed to be disambiguated, as it could be a part of an abbreviation, but was often used without a clear meaning, to observe in Figure 3.9 (3). If the annotation guidelines left problems unanswered, a link to a document was provided to resolve annotation issues.

Annotation Campaign

The annotation campaign was launched with four annotators, including three annotators working on French articles and two working on German (one annotator for both languages). An annotation planning specified the distribution of the workload, while the projects for the IAA needed to be annotated by everyone (in the respective language). The annotation campaign lasted for around four weeks.

During the annotation campaign, the questions which arose mostly revolved around the use of the tag *pers.ind.articleauthor*, i.e. when to use it and how to set the token boundaries. At times, unknown agencies and their distinction with authors also posed problems, and it was observed that most articles sampled from the 19th century were false positives, thus very few agencies were found for this time period.

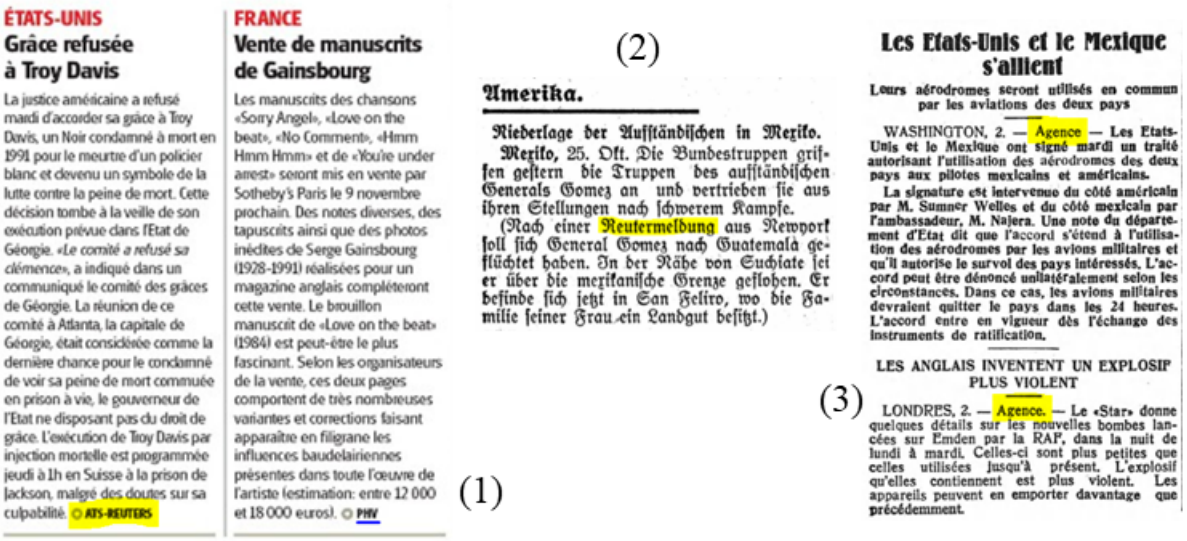


FIGURE 3.9
Examples for articles which might pose difficulties during annotation, taken from (1) *L'Impartial* 21/09/2011, (2) *Luxemburger Wort* 26/10/1927, and (3) *L'Impartial* 02/04/1941.

3.4.3 Inter-Annotator Agreement

To check for the Inter-Annotator Agreement (IAA) between the work of different annotators, we used a corpus of 48 articles per language, three from each decade. Inception (version 27.6) provided the metric Krippendorff's α (Krippendorff 2004), which is a measure from 0 to 1, with 0 reflecting no agreement, and 1 perfect agreement. According to Krippendorff (Krippendorff 2004, p. 241), a value of $\alpha \geq 0.8$ should be ideally met, although $\alpha \geq 0.667$ is still acceptable.

TABLE 3.2
Inter-annotator agreement (Krippendorff's Alpha) for the different tasks, languages and annotators (A1-A4) respectively.

	Annotators	Agency Tag	Transcript	Discard (Noisy OCR)
FR	**A1 - A2**	0.90	0.96	0.90
	A2 - A3	0.85	0.38	0.51
	A3 - A1	0.81	0.34	0.58
DE	**A1 - A4**	0.51	0.10	0.94

For the most important task, the tagging of agency mentions, all annotators for the French corpus were above the threshold of 0.8, see Table 3.2. The slight differences between the annotators are mostly due to missed annotations, normally either in the middle or at the end of the text. However, for the other two tasks, the transcription of corrected agency mentions and the discarding of articles due to too noisy OCR, the values are only good for the IA agreement between annotators A1 and A2. For the Transcript, the problem mainly comes from a different usage of the tag *pers.ind.articleauthor*. In the IAA corpus, not many agency mentions were noisy, leaving the article authors as the main target to correct. Because some annotators made use of the tag *pers.ind.articleauthor* more freely than others, they consequently also had to make more transcriptions, which explains the differences in Krippendorff's α. Similarly, not many articles in the IA corpus needed to be discarded due to noisy OCR, so different tags for two or three articles already caused the measure to decrease significantly.

For the German IA agreement, although the very low Transcript value of 0.1 might be justified (at least

- EXP-1938-03-15-a-i0108 (796): JTie note pessimiste d ' **Havas** PARIS , 15 . —
- EXP-1938-03-15-a-i0108 (810): Londres à l ' agence **Havas** : La déclaration du premier
- EXP-1956-05-07-a-i0124 (221): Willy Wittwer , 167 ; **AP** . Pierre Verron , 164
- EXP-1977-04-05-a-i0262 (574): jarrier (Fr) , **ATS** - Penske , à un
- EXP-1978-08-26-a-i0357 (310): , Ali . Ela . **Ap** . - 8 . Bloc

FIGURE 3.10
Excerpts of articles that contain an agency search term, but where the search term has not been annotated.

partly) with the same explanation as for the French counterpart, even the value for the agency tag is below the threshold of 0.667 (see Table 3.2). An error analysis shows that this is due to missed annotations, mistakenly annotated *ATS-SDA* as *unk*, differences in token boundaries (including a period or not) as well as a diverging use of the tag *pers.ind.articleauthor*. In general, the problems observed in the IA corpus were similar to the difficulties annotators reported in the document to resolve annotation issues during the annotation campaign.

While the tag *pers.ind.articleauthor* was not important for the continuation of the project, missed or wrongly annotated agencies had a more detrimental impact on the dataset's quality. Hence, we decided to include a post-correction step.

3.4.4 Annotation Post-processing

The annotation post-processing included a post-correction and a data enrichment step. As previously discussed, the IA agreement revealed several annotation difficulties, of which the wrongly tagged and missed agency mentions were of importance for the later classification process. Thus, the post-correction focused on those cases.

Checking *unk* Tags

The wrongly tagged agency mentions mainly concerned the *unk* tag. A correction of these was easily carried out by printing all occurrences and then correcting them manually. Through this process, mentions were transformed from *unk* to *ag, Havas, TASS, ATS-SDA* and *Extel*. Furthermore, the agency *Katholische Internationale Presseagentur (Kipa)* and the Chinese agency *Xinhua (Chine Nouvelle)* received their own classification tag, as they appeared several times in the training corpus.

Searching for Missed Annotations

To find missed annotations, we combined all search words used during the querying of the raw corpus collections in *impresso* (see Table B.2) into one regex query. With the help of this query, we retrieved all occurrences of potential agency mentions in the annotated dataset which were not annotated and printed them together with a window of 10 tokens around the word. For an example of the output, see Figure 3.10. Articles which seemed to have a missed annotation were then re-annotated in Inception.

In the end, 14 agency mentions were added in the German corpus and 71 mentions in the French corpus, amounting to 2.43% and 5.08% of all mentions in the German and French corpus respectively.

Enriching Annotations

As a last post-processing step, we added the article-level annotation whether an article contained a news agency, as well as the Wikidata item identifiers of the agencies and a coarse-grained token-level

annotation to specify if a token represented an agency or not. All of these additional annotations could be inferred from the token-level annotations and were executed with the help of search and insert/replace functionalities in Python. For example, in the case of Wikidata, a dictionary of agencies and their respective Wikidata identifier was manually assembled. Then, the respective identifier was added to each agency tag.

3.5 Final Dataset

The final dataset comprises six data files, providing a train, dev and test set for the French and German languages. The following section describes the characteristics of the final dataset, continues with the format of the data, and concludes with a short discussion of the strengths and limitations of the dataset.

3.5.1 Dataset Characteristics

The dataset contains 1,133 French and 397 German annotated documents, with 1,058,449 tokens, of which 1,976 have annotations. A breakdown into train, dev and test statistics can be found in Table 3.3. The difference in the number of agency mentions per set is due to the split being made on the document level and not on the basis of the number of mentions within the document. Thus, the German dev set features less than half of the mentions than the German test set. For the French corpus, the agency mentions are also more frequent in the test set, although they have more OCR noise (7%) than in the dev set (3%). The train set seems to be more balanced, with 5% noisy mentions of 1,113 annotations in total. The OCR noise for the German set is higher than the French, with 9% in the train and 8% in the test set. However, the test set is an exception, as only 3% of the annotated agency mentions are noisy.

TABLE 3.3
Overview of corpus statistics. %noisy gives the percentage of agency mentions with at least one OCR error.

	Lg.	Docs	Tokens	Mentions	%noisy
Train	de	333	247,793	493	9%
	fr	903	606,671	1,122	5%
Total		1,236	854,464	1,615	6%
Dev	de	32	28,745	26	8%
	fr	110	77,746	114	3%
Total		142	106,491	140	4%
Test	de	32	22,437	58	3%
	fr	120	75,057	163	7%
Total		152	97,494	221	6%
All	de	397	298,975	577	9%
	fr	1,133	759,474	1,399	5%
Total		1,530	1,058,449	1,976	6%

The slight deviation from the 80/10/10 train/dev/test set ratio is due to the articles which were completely discarded because of bad OCR quality. In total, this concerned 15 German and 65 French articles, the majority of discarded articles being from the 19th century, as can be seen in Figure 3.11.

Figure 3.12 shows the number of agency mentions within the corpus split over decades. It reveals that in the German corpus, agencies were cited only in the 20th century, while the mentions of agencies in the French corpus already start in 1850 and increase continuously (with some disruptions) over time.

A comparison of agency appearances in the French and German corpora (Figures 3.13 and 3.14) exposes

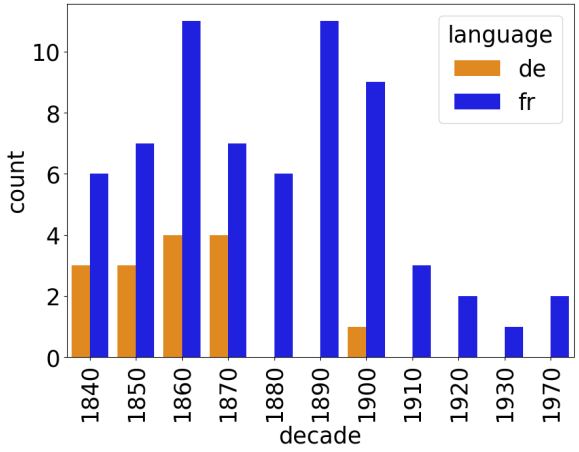


FIGURE 3.11
Number of discarded articles due to bad OCR quality, per language and decade.

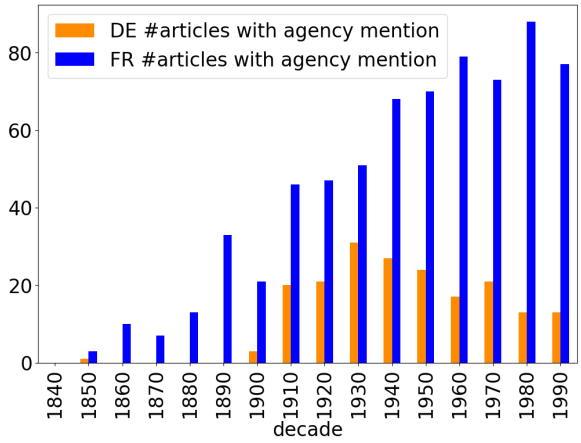


FIGURE 3.12
Number of articles which include at least one agency mention, per language and decade.

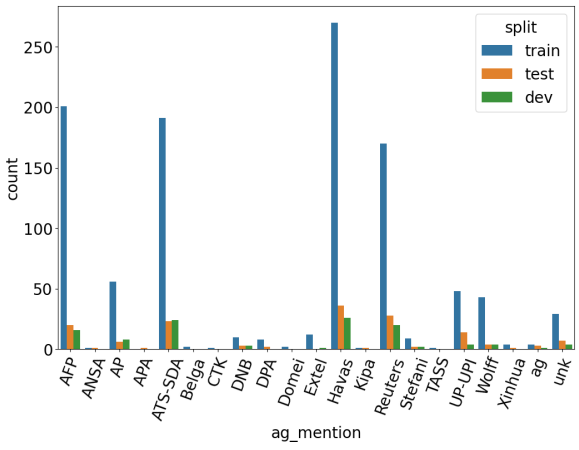


FIGURE 3.13
Number of occurrences of news agencies in the train/dev/test split for the French corpus.

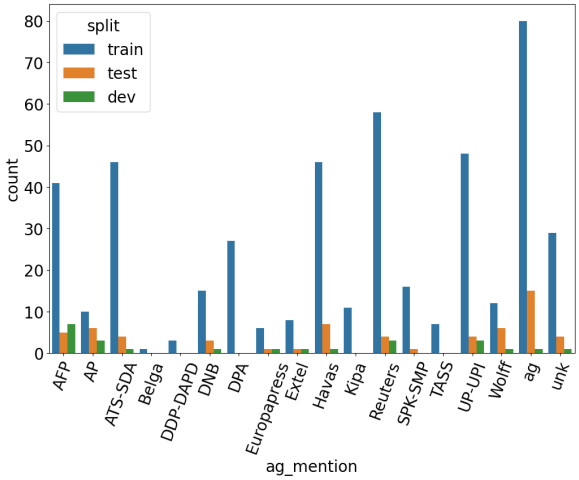


FIGURE 3.14
Number of occurrences of news agencies in the train/dev/test split for the German corpus.

```
# global.columns = TOKEN NE-COARSE-LIT NE-COARSE-METO NE-FINE-LIT NE-FINE-METO NE-FINE-COMP NE-NESTED NEL-LIT NEL-METO RENDER SEG OCR-INFO MISC
# language = fr
# newspaper = EXP
# date = 1924-03-27
# document_id = EXP-1924-03-27-a-i0077
# news-agency-as-source = Q2826560
# segment_iiif_link = https://impresso-project.ch/api/proxy/iiif/EXP-1924-03-27-a-p0005/224,107,285,87/full/0/default.jpg
POLITIQUE       O       O       O       O       O       O       _       _       EndOfLine       _       _       _
# segment_iiif_link = https://impresso-project.ch/api/proxy/iiif/EXP-1924-03-27-a-p0005/160,202,398,53/full/0/default.jpg
France  O       O       O       O       O       O       _       _       _       _       _       _
et      O       O       O       O       O       O       _       _       _       _       _       _
Grande  O       O       O       O       O       O       _       _       NoSpaceAfter    _       _       _
-       O       O       O       O       O       O       _       _       NoSpaceAfter    _       _       _
Bretagne        O       O       O       O       O       O       _       _       EndOfLine       _       _       _
# segment_iiif_link = https://impresso-project.ch/api/proxy/iiif/EXP-1924-03-27-a-p0005/200,239,319,52/full/0/default.jpg
Une     O       O       O       O       O       O       _       _       _       _       _       _
invention       O       O       O       O       O       O       _       _       _       _       _       _
et      O       O       O       O       O       O       _       _       _       _       _       _
un      O       O       O       O       O       O       _       _       _       _       _       _
démenti O       O       O       O       O       O       _       _       EndOfLine       _       _       _
# segment_iiif_link = https://impresso-project.ch/api/proxy/iiif/EXP-1924-03-27-a-p0005/129,269,488,53/full/0/default.jpg
LONDRES O       O       O       O       O       O       _       _       NoSpaceAfter    _       _       _
,       O       O       O       O       O       O       _       _       _       _       _       _
27      O       O       O       O       O       O       _       _       _       _       _       _
(       O       O       O       O       O       O       _       _       NoSpaceAfter    _       _       _
Havaa   B-org   O       B-org.ent.pressagency.Havas     O       O       O       Q2826560        _       NoSpaceAfter    _       Transcript:Havas|LED0.20        _
)       O       O       O       O       O       O       _       _       NoSpaceAfter    _       _       _
.       O       O       O       O       O       O       _       _       _       PySBDSegment    _       _
—       O       O       O       O       O       O       _       _       _       _       _       _
Le      O       O       O       O       O       O       _       _       _       _       _       _
<       O       O       O       O       O       O       _       _       _       _       _       _
Daily   O       O       O       O       O       O       _       _       EndOfLine       _       _       _
# segment_iiif_link = https://impresso-project.ch/api/proxy/iiif/EXP-1924-03-27-a-p0005/106,292,511,53/full/0/default.jpg
Herald  O       O       O       O       O       O       _       _       NoSpaceAfter    _       _       _
-       O       O       O       O       O       O       _       _       NoSpaceAfter    _       _       _
a       O       O       O       O       O       O       _       _       _       _       _       _
```

FIGURE 3.15
Beginning of an article in IOB format (tab-separated values).

differences as well. For the French language, mainly the four agencies AFP, ATS, Havas and Reuters are present, while the agencies in the German newspapers are more evenly distributed, with DPA, UP-UPI, the generic *ag* and agencies labelled as *unk* having a similar frequency as the four aforementioned agencies. The agency distributions overtrain, dev and test split are mostly as planned. Some agencies are finally not present in all three sets, i.e. Belga, Extel or TASS in French and DDP, DPA or TASS in German. However, they are always present in the respective training set, which should ensure stable training of the classifier.

3.5.2 Final Format

For the final format, we adopted the schema of HIPE-2022 (Ehrmann, Romanello, Najem-Meyer et al. 2022), which in turn is based on the CoNLL-U format[2], a tab-separated column textual format using an IOB tagging scheme (inside-outside-beginning format): It saves each token of an article in a new row, and all the information about this token is added in the same row, organized as tab-separated columns. An example of the format can be found in Figure 3.15.

To get to this format, we first performed a tokenization. This included the separation of apostrophes and hyphens from a token into two tokens, and the substitution of unprintable characters with the underscore. Secondly, we used the sentence segmentation algorithm by the PySBD library (Sadvilkar and Neumann 2020) to split the text into sentences. Information on those steps was saved in the columns `Render` and `Seg` (in the column textual format) respectively.

The annotations on the agency mentions can be found in multiple columns: `NE-Coarse` only states if the annotated token is an organisation or a person, while `NE-Fine` also specifies the news agency (e.g. *org.ent.pressagency.AFP*) or displays the tag *pers.ind.articleauthor*. Both granularity levels include the prefix *B-* or *I-*, indicating if the annotation began with the current token or if it is inside (e.g. *Press* of United *Press* would have the prefix *I-*). Furthermore, the Wikidata IDs are saved in the `NEL` (named entity linking) column, while the `OCR-Info` provides the manual transcript of the agency (if any), as well as the Levenshtein distance (LED) between the original token and its correct transcript (LED= 0 if no transcript is given).

[2] `https://universaldependencies.org/format.html`

Each data file, one for train, dev and test in the two languages respectively, begins with the information on the column names used for the document. Further metadata is written on top of each article, including the language, newspaper, publication date and document ID of the article. The line *news agency as source* is the article-level tag specifying if a news agency is mentioned as a source within this article. This information is presented through the corresponding Wiki-ID of the mentioned news agency. As an example, see Figure 3.15; the highlighted parts show the format of a typical annotation for this project.

Finally, IIIF[3] links for each text segment in the original newspaper article were added, referring to an image with the facsimile of the corresponding segment.

3.5.3 Strengths and Limitations

During the building of the dataset, we proceeded as circumspectly as possible. The selection of news agencies was based on a literature search and queries in the *impresso* corpus, and decisions like the application of thresholds were motivated through interactions with the data. Thus, the dataset can be assumed to contain a representative set of news agencies cited in Swiss and Luxembourgish newspapers in the 19th and 20th centuries. Additionally, the manual process of annotation as well as the subsequent addition of missed annotations ensured a good quality of the agency annotations.

One of the main limitations of the dataset is its reliance on the agency mentions to identify an article as a news agency article. The grey area of news agency articles which do not contain an attribution to an agency could not be included; and although we tried to design the queries for news agencies as thoroughly as possible, we might have missed some relevant data during the construction of the raw corpus.

When using the dataset, one should be aware that German articles are underrepresented compared to the French data, which could make the training in the German language less stable. Other imbalances include the different levels of OCR noise in the train, dev and test sets, as well as a skew of the annotations to the 20th compared to the 19th century. Additionally, the tag *pers.ind.articleauthor* should be treated with caution, as it was annotated on the side and hence might occur inconsistently.

Keeping those limitations in mind, the dataset provides valuable knowledge and should be sufficient to train a good classifier, which is the content of the following chapter.

[3] `https://iiif.io/api/image/3.0/`

4 Experiments

After constructing and annotating the dataset, we conducted a series of experiments to learn efficient models for the tasks of news agency mention recognition (NER) and text classification of articles containing agency content in French and German. This chapter introduces the experimental setup, presents the models and analyses their performance, and concludes with an evaluation of their strengths and limitations, which guided the choice of the two models (one for each language) used for the subsequent inference process on the *impresso* corpus.

4.1 Experimental Settings

We conducted a variety of experiments to find the best-performing model for our dataset, mostly varying model architectures (Section 4.1.1), along with an ablation study regarding the hyperparameter "maximum sequence length" (Section 4.1.2). For the evaluation of the results, we used the proper metrics as well as a lookup baseline, which are introduced in Section 4.1.3.

The implementation relies on the deep learning framework PyTorch (Paszke et al. 2019) and HuggingFace[1], which provides off-the-shelf pre-trained language models[2]. The training of our models for NER and text classification took place on two GPUs (NVIDIA Corporation GM200) with a memory of 12 GB each.

4.1.1 Model Architectures and Specifications

As the base model architecture, we chose BERT (Devlin et al. 2019), because it is stable, widely used (see Section 2.2) and exists in many versions trained on various languages, including in-domain (historical newspapers) data[3]. As part of its pre-training procedure, BERT uses a pair of sentences for two tasks: masked language model (MLM) and next sentence prediction (NSP). MLM is performed by the replacement of a percentage of the input tokens with a [MASK] token. The model then attempts to predict the original value of the masked words, based on the context provided by the other, non-masked, tokens in the sentence. During NSP, the model receives pairs of sentences as input and learns to predict if the second sentence in the pair is the subsequent sentence in the original document.

Moreover, BERT can be fine-tuned for different downstream tasks, i.e. for both named entity recognition and text classification, the two tasks needed for this project. Indeed, we decided on a multitask approach (S. Chen, Zhang and Yang 2021), meaning that the same model is fine-tuned on both tasks at the same time. For each task, a separate feed-forward layer is added on top of BERT, creating the two different outputs, i.e. a news agency or "*O*" on token-level and a *yes/no* (0/1) for classification of agency content. Then, the sum of the respective losses is used for backpropagation.

[1] `https://huggingface.co/`

[2] The code for the experiments in this chapter can be found on Github under `https://github.com/impresso/newsagency-classification/tree/main/lib/bert_classification`.

[3] `https://huggingface.co/docs/transformers/model_doc/bert`

TABLE 4.1
Specifications of the different models tested. The column *Language* indicates the language of the dataset used for fine-tuning and evaluation. "multi" signifies fine-tuning both in French and German, with the evaluation done in the two languages separately.

HuggingFace Model	**Reference**	**Base Model**	**#params**	**Training Corpus**	**Language**
bert-base-cased	Devlin et al. 2019	BERT	110M	BooksCorpus[4] (800M words), English Wikipedia (2,500M words)	fr, de, multi
bert-base-multilingual-cased	Devlin et al. 2019	BERT	110M	top 104 languages with the largest Wikipedia[5]	fr, de, multi
xlm-roberta-base	Conneau, Khandelwal et al. 2019	RoBERTa	270M	2.5TB of filtered CommonCrawl data containing 100 languages	fr, de, multi
camembert-base	Martin et al. 2020	RoBERTa	110M	French part of the OSCAR corpus[6], 138GB text (32.7B tokens)	fr
bert-base-german-cased	Chan et al. 2019	BERT	110M	Wikipedia (6GB), OpenLegalData (2.4 GB), news articles (3.6 GB)	de
bert-base-historic-multilingual-cased	Schweter, März et al. 2022	BERT	110M	196B subtokens/32K vocabulary size (French, German, Finnish & Swedish Europeana newspapers[7]; British Library[8])	fr, de, multi
bert-base-french-europeana-cased	Schweter 2020	BERT	110M	French Europeana newspapers(63GB/11B tokens)	fr
bert-base-german-europeana-cased	Schweter 2020	BERT	110M	German Europeana newspapers(51GB/8B tokens)	de

BERT is limited by a maximum sequence length of 512 tokens. This means that in order to classify articles on their possible news agency content, they have to be split up into different segments. We chose to split on sentence level, due to the fact that the training process of BERT-based models uses sentence-level context and sentences are generally self-contained units of meaning. A sentence was classified as having agency content if it featured a news agency mention. The article classification could then be inferred from the collective information of its sentences.

Table 4.1 lists the different model architectures which were tested during the experiments. Most of the models are based on the original BERT architecture, which consists of 12 transformer blocks, a hidden vector size of 768 and a multi-head attention with 12 self-attention heads. Two out of the eight models that we experimented with use RoBERTa. The difference between RoBERTa to BERT does not lie in its model architecture, but in the training process, i.e. RoBERTa removes the NSP task based on the observation that it does not contribute significantly to the model's performance (and could even hurt it Liu et al. 2019).

[4]BooksCorpus (Zhu et al. 2015)
[5]No further specification was given.
[6]OSCAR Suarez 2019, a pre-filtered and pre-classified version of Common Crawl
[7]`http://www.europeana-newspapers.eu/`
[8]Digitised Books by British Library Labs 2016

As the chosen models for this project were not pre-trained according to the original RoBERTa paper, the adjusted hyperparameters remain the only distinction to the BERT setup. Except for the 270M parameter xlm-roberta-base, all models rely on the BERT-based model with 110M. We selected the cased versions of the respective models because casing makes a difference for the detection of news agencies (e.g. "AP" vs. "ap").

The main difference between the tested models can be found in their (pre-)training corpus. The models *bert-base-cased, bert-base-multilingual-cased, xlm-roberta-based, camembert-base* and *bert-base-german-cased* were trained on general corpora, of which some contained texts in a specific language (*camembert, bert-base-german, bert-base*), while others featured multiple languages (*bert-base-multilingual, xlm-roberta-base*). The same categorization can be made for the remaining three models *bert-base-historic-multilingual-cased, bert-base-french-europeana-cased* and *bert-base-german-europeana-cased* (the language being included in their model descriptor), with the distinction that they were pre-trained on in-domain data, specifically parts of the Europeana newspaper corpus, which is a collection of historical European newspapers ranging from the 18th to the 20th century (Neudecker and Antonacopoulos 2016; Schweter, März et al. 2022). The descriptions on the training corpora in Table 4.1 only give a rough idea about the nature of the texts used for the pre-training; for a closer view of the preparation of the data (cleaning, filtering, tokenization) we refer to the literature (see the Reference column of Table 4.1). However, one has to keep in mind that these details can make a considerable difference in the performance of the models (Liu et al. 2019).

Regarding fine-tuning, the models were either trained on the French or German news agency datasets, or both (multilingual training). The languages for the respective models were selected based on their pre-training data. This information is displayed in the *Language* column of Table 4.1. For the multilingual option, the evaluation was performed on the German and French data separately.

4.1.2 Hyperparameters and Ablation Studies

Most hyperparameters were set according to the recommendations of Devlin et al. 2019 and remained fixed during the experiments. Thus, the chosen hyperparameters for the experiments are the following:

- Maximum sequence length: 64, 128, 256 and 512;
- Batch size: 16 (except with the maximum sequence length of 512 where the batch size was 8);
- Learning rate: 5×10^{-5} (10% of training steps used for the learning rate warm-up);
- Epochs: 3;
- Optimizer: AdamW (Loshchilov and Hutter 2019) with $\epsilon = 1e - 8$ and L_2 weight decay of 0);
- Dropout: 0.1 (all layers);
- Layer activation function: Gaussian error linear unit (GELU) (Hendrycks and Gimpel 2016);
- Loss: cross-entropy loss.

Regarding the maximum sequence length, the impact of its variation depends on the nature of the underlying training corpus. The sequence length is generally set to 64, 128, 256 or 512, smaller lengths reducing the computation cost of the model. If the input sequence is longer, it is cut at the maximum sequence length. If it is shorter, the input is padded with a special token depending on the pre-trained model (e.g. [PAD]).

The news agency datasets for this project exhibit an average sentence length of 25 for French and 24 for German. For French, 98% of all sentences are shorter than 98 tokens, while 99% of all sentences are shorter than 166 tokens. Those numbers are higher for the German part of the corpus, with 98% of all

sentences being shorter than 145 tokens and 99% being shorter than 225 tokens. This suggests that a lower maximum sequence length might already be sufficient to capture most of the data. Still, when clipping a sentence at a certain length, there is the risk that agency mentions are removed as well. Table 4.2 gives statistics on this phenomenon. For a maximum sequence length of 128, for example, 85 agency mentions would be discarded in the French and 34 in the German corpus, amounting to around 6% and 7% of all mentions in the respective languages.

TABLE 4.2
Number of discarded news agency mentions in the datasets according to the maximum sequence length.

Max. Sequence Length		**128**	**256**	**512**
Missed	**FR**	85	54	21
	DE	34	14	2

To investigate the relationship between the maximum sequence length and classification performance and to choose the optimal value, we systematically examined this hyperparameter in an ablation study. We had to halve the batch size for the maximum sequence length of 512 due to limited GPU capacity.

4.1.3 Evaluation Methodology

For the evaluation, we used the metrics *precision*, *recall* and the *F-score*. They use the categorization of classified items into true positives (TP), false positives (FP) and false negatives (FN)[9]. Regarding, for example, the sentence classification for agency content, TP are those sentences correctly classified as having agency content, while FP are those falsely attributed to agencies, and FN are the sentences "missed" by the classifier, i.e. falsely classified as not having agency content.

Precision and recall are defined as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

and

$$\text{Recall} = \frac{TP}{TP + FN}$$

The F-score (or *F1*) is the harmonic mean of precision and recall, i.e.:

$$\text{F-Score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

For more than two classes, the metrics need to be averaged. One possibility would be to compute the scores separately for each class and then average over them (macro scores). We chose to work with the micro scores, which sum all individual TP, TN and FN counts, before computing the respective metric. As an example, the micro precision would be computed through:

$$\text{Micro Precision} = \frac{\sum_{i=1}^{n} TP_i}{\left(\sum_{i=1}^{n} TP_i\right) + \left(\sum_{i=1}^{n} FP_i\right)},$$

where n signifies the number of classes and TP_i (FP_i) the number of TP (FP) in class i.

[9]The fourth category is true negatives, but they are irrelevant here.

Additional to the in-model evaluation scorer which took the input of the model as a basis, we consulted the HIPE-scorer from the HIPE campaign[10] for the task of news agency recognition. This made it possible to incorporate those agency mentions that were discarded by the model due to its dependence on the maximum sequence length, and to get evaluations split by time and OCR noise level.

4.1.4 Lookup Baseline

In order to have a baseline against which we could compare our deep-learning classification results, we constructed a lookup baseline for the NER task. It built on the agency searchwords which had been used for the collection of the raw corpus in *impresso* (Appendix B.2), discarding the 18 queries with more than one word for simplicity reasons. Every token in the dev and test sets which matched one of the remaining 156 agency searchwords, disregarding casing, was classified with the associated news agency. The results are rather high, as can be seen in Table 4.3. The significantly higher recall than precision can be traced back to two factors: Firstly, the searchwords were constructed as comprehensively as possible, and secondly, a simple lookup finds many false positives. For instance, the searchword *AP* matches with any *ap/Ap/AP/aP* token, also those which were originally part of a longer word, e.g. OCR errors like "*Ap* penzell" or splits due to line breaks like "*ap* - peler".

TABLE 4.3
Evaluation results for the NER lookup baseline (micro average).

		Precision	Recall	F1
FR	**dev**	0.57	0.85	0.68
	test	0.67	0.80	0.73
DE	**dev**	0.50	0.75	0.60
	test	0.77	0.87	0.82

4.2 Results

For the presentation of the results, we first compare the performances of the different model configurations for both the news agency entity recognition and sentence classification tasks, before providing a general summary (Section 4.2.1). We then examine in more detail the strengths and weaknesses of the models from different perspectives, such as class distribution, OCR noise, or performance over time (Section 4.2.2).

4.2.1 Model Performance

Every model configuration was run five times with different seeds to get a more reliable picture of their performance. The mean and standard deviation of the five runs are the basis of this section. Results can be found in Table 4.4 and Table 4.5 for news agency entity recognition and sentence classification, respectively, for a maximum sequence length of 128. We set this parameter to a fixed value for reasons of comprehensibility, and 128 is a fitting representative for all maximum sequence lengths. Figures 4.1 and 4.2 visualize the results for the French and German entity recognition experiments. Similar graphs for the sentence classification and additional material such as the full tables with all maximum sequence lengths or results on the dev set are in Appendix E.

The results presented in this section are based on in-model evaluations, i.e. by an evaluation function run on the input of each model. A comparison of the NER task with the HIPE scorer (see Section 4.1.3) showed that, against our expectations, the HIPE results are generally better than the in-model evaluations (0.02-0.03 on average) and have a tendency to be better with higher maximum sequence length (for more

[10] https://github.com/hipe-eval/HIPE-scorer

TABLE 4.4
Results for named entity (agency) recognition for German and French test sets, for models with maximum sequence length equal to 128. Experiments were run five times per configuration, the values show the mean and the standard deviation in brackets.

Language (train-test)	Model	F1	Precision	Recall
fr-fr	bert_base_cased	0.728 (0.022)	0.723 (0.025)	0.735 (0.022)
	bert_base_multilingual_cased	0.607 (0.130)	0.600 (0.123)	0.614 (0.137)
	camembert_base	0.640 (0.025)	0.692 (0.040)	0.595 (0.023)
	dbmdz_bert_base_french_europeana_cased	**0.781 (0.025)**	**0.774 (0.046)**	**0.790 (0.010)**
	dbmdz_bert_base_historic_multilingual_cased	0.705 (0.043)	0.689 (0.054)	0.723 (0.033)
	xlm_roberta_base	0.655 (0.018)	0.780 (0.024)	0.566 (0.028)
multilingual-fr	bert_base_cased	**0.770 (0.019)**	**0.766 (0.018)**	**0.774 (0.023)**
	bert_base_multilingual_cased	0.717 (0.069)	0.710 (0.078)	0.724 (0.060)
	dbmdz_bert_base_historic_multilingual_cased	0.684 (0.080)	0.658 (0.081)	0.714 (0.079)
	xlm_roberta_base	0.655 (0.023)	0.729 (0.031)	0.595 (0.023)
	Lookup Baseline	0.729	0.669	0.802
de-de	bert_base_cased	**0.838 (0.027)**	**0.827 (0.045)**	**0.851 (0.011)**
	bert_base_german_cased	0.830 (0.029)	0.819 (0.051)	0.842 (0.017)
	bert_base_multilingual_cased	0.093 (0.119)	0.100 (0.112)	0.091 (0.123)
	dbmdz_bert_base_german_europeana_cased	0.311 (0.227)	0.309 (0.227)	0.313 (0.227)
	dbmdz_bert_base_historic_multilingual_cased	0.374 (0.264)	0.356 (0.266)	0.396 (0.262)
	xlm_roberta_base	0.727 (0.026)	0.760 (0.045)	0.698 (0.027)
multilingual-de	bert_base_cased	**0.828 (0.018)**	**0.811 (0.039)**	**0.847 (0.009)**
	bert_base_multilingual_cased	0.748 (0.053)	0.718 (0.046)	0.781 (0.063)
	dbmdz_bert_base_historic_multilingual_cased	0.664 (0.172)	0.622 (0.175)	0.713 (0.165)
	xlm_roberta_base	0.776 (0.009)	0.790 (0.029)	0.762 (0.01)
	Lookup Baseline	0.817	0.770	0.870

details see Appendix E.1). Keeping this in mind, the in-model evaluations still seem to be a solid choice for a discussion on model performances.

News Agency Entity Recognition. The results for the Entity Recognition task presented in Table 4.4 are relatively high, with F-scores of 0.781 in French and 0.838 in German – especially compared to historical NER, which usually achieves F-scores in the order of 0.6-0.7(Ehrmann, Hamdi et al. 2023). However, the underlying task is also easier, as the models are only required to detect a finite, rather low number of named entities. The *unk* is an exception, as the named entities which are subsumed in this class can come from many different, previously unseen news agencies. However, its results are not weighted heavily, as it is only one class.

Comparing the French and German results, they are generally lower for French, which could be due to the higher ratio of OCR noise in the French test set (7%) than in the German test set (3%). The error

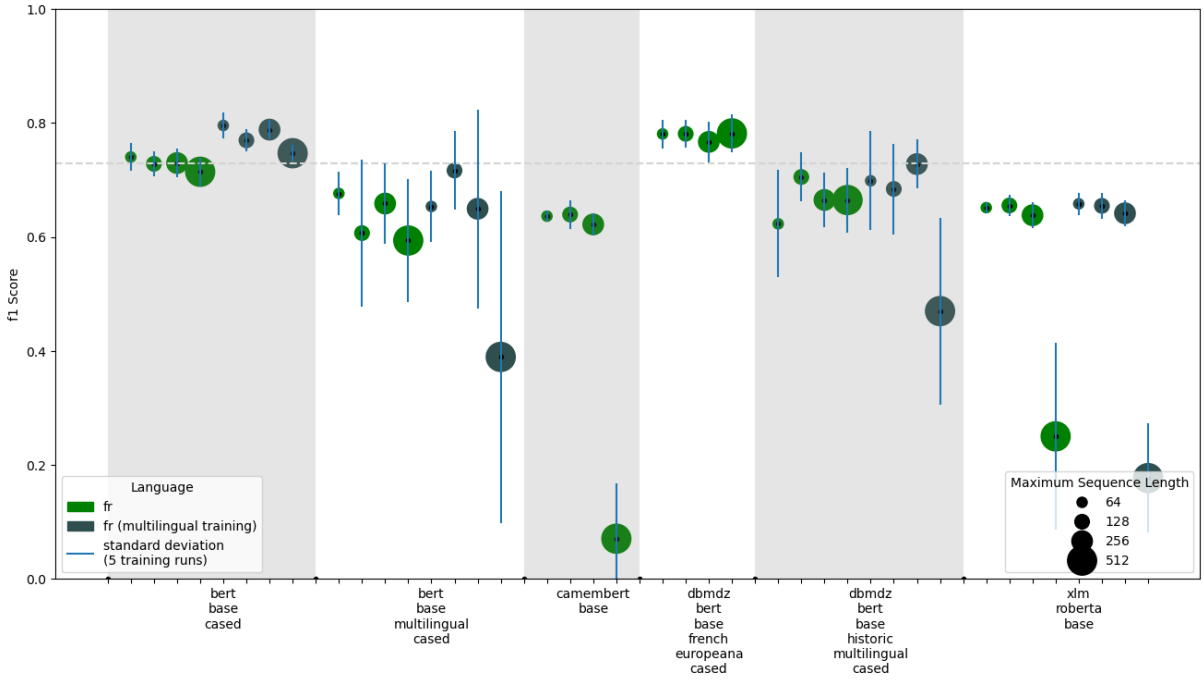


FIGURE 4.1

Results for agency recognition on the French test set. Experiments were run five times per configuration, the dots present the mean, the blue lines the standard deviation. The colour of the dots refers to the training set (French or Multilingual), while the size specifies the maximum sequence length. The lookup baseline is displayed as a grey dashed line (see Section 4.1.3).

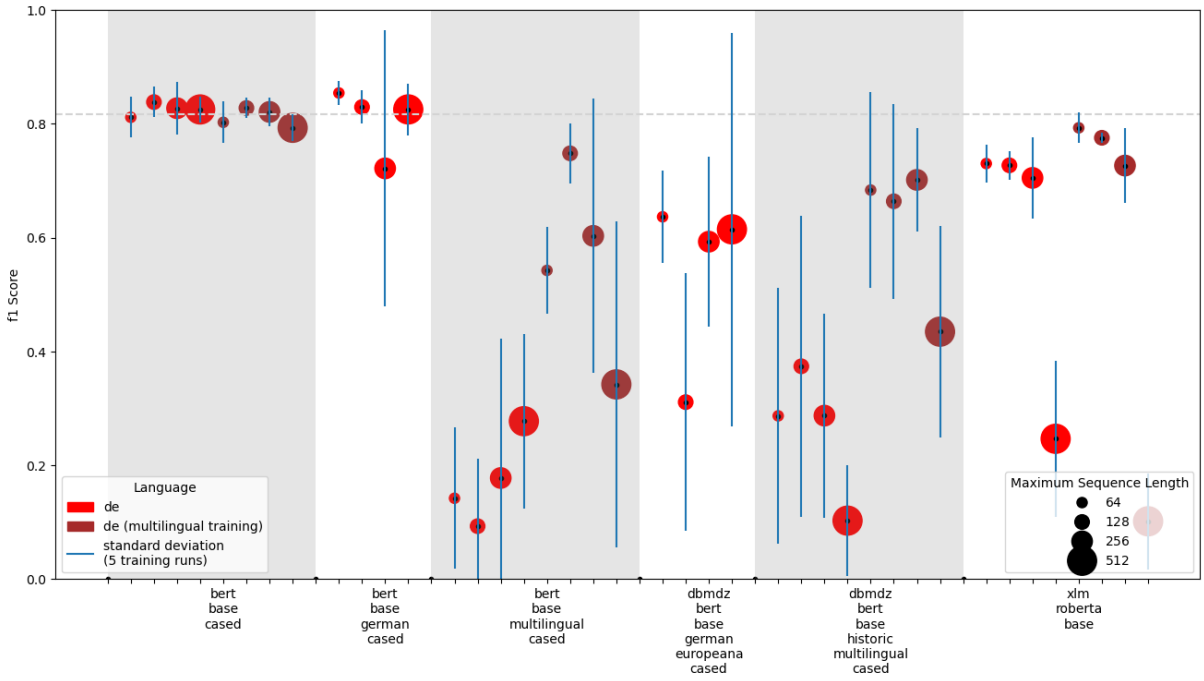


FIGURE 4.2

Results for agency recognition on the German test set; the layout specifications are the same as for Figure 4.1.

analysis in Section 4.2.2 examines this argument further. Looking at Figure 4.2, it is striking that the standard deviations for the German results are a lot higher than the French. This can be traced back to the lesser amount of training data for the German language, making training less stable (Devlin et al. 2019). The best model for French is *bert-base-french-europeana-cased* with an F-score of 0.781, a precision of 0.774 and a recall of 0.790. Considering the models trained in both French and German, *bert-base-cased* can compete, achieving an F-score of 0.770. For German, surprisingly *bert-base-cased* fine-tuned on the German dataset is the best model, with 0.838 F-score, 0.827 precision and 0.851 recall. Its version fine-tuned in both languages attains the highest values for the multilingually fine-tuned models (0.828 F-score), although the second highest F-score, in general, is acquired by *bert-base-german-cased* trained on German data (0.830 F-score). Why does *bert-base-cased*, a model pre-trained on English Wikipedia and an English Book Corpus (see 4.1), perform so well on German (and also French) data? One supposition would be that the languages exhibit similarities, especially English and German, which stem from the same language family. Thus, for a task which "only" needs to recognize named entities (proper nouns), it might be easy to transfer knowledge from English to German. Research on zero-shot transfer learning across languages (Deshpande, Talukdar and Narasimhan 2022, Wu and Dredze 2019) supports this hypothesis, showing that performances are higher if the languages share sub-words and have aligned word embeddings. Additionally, Wikipedia data is rich in named entities (e.g. Nothman, Curran and Murphy 2008) and might contain quotes from other languages, which also might help the news agency recognition task at hand. On the technical side, the hyperparameters chosen for the fine-tuning were oriented on the original (English) BERT implementation, so it is possible that they also contributed to high and stable results for *bert-base-cased*.

Compared to the lookup baseline, many models underperform – at least in terms of F-score. When breaking the results down to precision and recall, the models generally perform better than the baseline for precision, but miss more mentions than the lookup algorithm, which results in lower recall. Following the assessment of a consulted historian, it is easier to work with a "cleaner" corpus which misses a few articles than to get wrong research results because of too many false positives. Thus, higher precision is valuable, which shows that news agency classification with a simple lookup algorithm is not enough to get satisfying results.

Regarding the differences between models, some models provide unstable results (see Figures 4.1 and 4.2). For French and German, *bert-base-multilingual-cased* shows a lot of variation, both between different runs of the same model configuration and between different maximum sequence lengths. A reason for this might be an insufficient amount of pre-training data per language (Conneau and Lample 2019). The same argumentation could apply to the models trained on in-domain (Europeana) data, where the German results vacillate a lot. Although the amount of pre-training data for the French Europeana model is similar to the German Europeana model, the French model could be fine-tuned on more news agency data, which could explain its comparatively stable behaviour.

Another observation within our results was the occasional decrease in performance when the maximum sequence length was set to 512 tokens. This could be attributed to the increase in resource requirements associated with this lengthier sequence, which necessitated the reduction of the batch size from 16 to 8 due to our GPU capacity constraints. When we experimented with a significantly smaller batch size of 2, the models failed to learn effectively, resulting in very low performance. Indeed, smaller batch sizes, while they use less memory and are computationally less demanding, can lead to noisier gradient estimates during training. This is because the error gradients calculated during backpropagation are based on fewer examples, which may not be as representative of the overall data distribution. Although the adjustment of other hyperparameters, such as learning rate or gradient accumulation steps, might have potentially mitigated these performance issues, the main objective of this study was to identify a single, high-performing model rather than pursuing comprehensive hyperparameter optimization.

TABLE 4.5
F-scores of sentence classification on French and German test sets, for models with a maximum sequence length of 128, considering all sentences and only the "positive" ones, i.e. those with an agency mention. Experiments were run five times per configuration, the values show the mean and the standard deviation in brackets.

Language (train-test)	Model	F1 (Both Classes)	F1 ("Positive" Class)
fr-fr	bert_base_cased	0.988 (0.000)	0.877 (0.005)
	bert_base_multilingual_cased	0.989 (0.001)	0.895 (0.010)
	camembert_base	0.983 (0.001)	0.830 (0.011)
	dbmdz_bert_base_french_europeana_cased	**0.990 (0.001)**	0.899 (0.013)
	dbmdz_bert_base_historic_multilingual_cased	**0.990 (0.001)**	**0.902 (0.011)**
	xlm_roberta_base	0.983 (0.001)	0.818 (0.016)
multilingual-fr	bert_base_cased	0.989 (0.001)	0.896 (0.012)
	bert_base_multilingual_cased	**0.990 (0.001)**	**0.899 (0.013)**
	dbmdz_bert_base_historic_multilingual_cased	**0.990 (0.001)**	**0.899 (0.009)**
	xlm_roberta_base	0.983 (0.001)	0.826 (0.008)
de-de	bert_base_cased	0.985 (0.002)	0.842 (0.02)
	bert_base_german_cased	**0.989 (0.003)**	**0.884 (0.028)**
	bert_base_multilingual_cased	0.986 (0.002)	0.858 (0.021)
	dbmdz_bert_base_german_europeana_cased	0.985 (0.002)	0.841 (0.015)
	dbmdz_bert_base_historic_multilingual_cased	0.986 (0.001)	0.860 (0.012)
	xlm_roberta_base	0.984 (0.002)	0.830 (0.018)
multilingual-de	bert_base_cased	0.985 (0.001)	0.843 (0.012)
	bert_base_multilingual_cased	0.985 (0.002)	0.850 (0.014)
	dbmdz_bert_base_historic_multilingual_cased	**0.987 (0.002)**	**0.865 (0.017)**
	xlm_roberta_base	0.986 (0.002)	0.853 (0.021)

Sentence Classification. Regarding sentence classification, the results are much less variable, both between the different models and between different maximum sequence lengths. We only see drops for the maximum sequence length for *bert-base-multilingual-cased*, *camembert-base* and *xlm-roberta-base* (see Appendix E.3), but they are seldom as extreme as for the entity recognition task. Moreover, the results generally exhibit a very high F-score (up to 99%, see Table 4.5), but this is mainly due to the high number of correctly classified "negative" sentences, i.e. those sentences which do not contain an agency mention. The F-scores for the "positive" class, i.e. the class of sentences with an agency mention, are therefore more informative.

In contrast to the entity recognition task, French and German results are similarly high. For French, three models perform equally well, namely *bert-base-french-europeana-cased* pre-trained on French historical newspaper data, *bert-base-multilingual-cased* pre-trained on multilingual data, and *bert-base-historic-multilingual-cased*; the latter, pre-trained on multilingual historical newspaper data, displays top results both for French only and for multilingual fine-tuning. Their general F-score lies at 0.990, while the F-scores of the positive class are nine percentage points lower, at 0.902 for the historic multilingual model and 0.899 for the rest. For German, *bert-base-german-cased* shows the best results with F-scores of 0.989 and 0.884 for both classes and the positive class respectively, but *bert-base-historic-multilingual-cased* with multilingual training can compete with a general F-score of 0.987 and a slightly lower F-score than for German BERT for the positive class of 0.865.

General Trends. Based on the literature, we expected models trained on in-domain data to perform better (Ehrmann, Hamdi et al. 2023), as well as those with a bigger model size (Devlin et al. 2019) and with more pre-training data and a higher maximum sequence length (Liu et al. 2019). These expectations were only partially met. While the in-domain models' performances were among the best, they did not always perform well for the entity recognition task, especially for German. Compared to e.g. XLM-RoBERTa with its 2.5 TB of data, the 51 GB of german-europeana might have not been enough, even though they came from in-domain corpora. Indeed, according to Ehrmann, Hamdi et al. 2023, "what is best between very large modern vs. in-domain LMs remains an open question". Still, not all generic models were pre-trained on more data than the Europeana models. German BERT, for example, was even trained on less. As well as quantity, the size of the underlying vocabulary and the quality of the data also play an important role in the performance of the models (Conneau, Khandelwal et al. 2019, Whang et al. 2023). Checking the years for which data was available in the Europeana corpora[11], one can see that apart from a few years, the French corpus has a more equal distribution of data across time than the German corpus, which might have influenced the higher performance of the French Europeana BERT model.

Similarly, a larger model size does not guarantee better results, as shown by the results of *XLM-RoBERTa*, the only model with a higher parameter number of parameters. It even did not outperform multilingual BERT like in the original paper by Conneau and Lample 2019. An explanation could be the missing hyperparameter search during the experiments, but this would need to be verified by a more comprehensive analysis. An observation which can be made for all models pre-trained on multilingual data is that they perform better on the entity recognition task when fine-tuned on both French and German. Schweter, März et al. 2022 found similar results for the evaluation of historic multilingual BERT on a NER task for German, although, for English and French, the models only fine-tuned on the respective languages performed better. The difference between single-language vs. multilingual fine-tuning is especially striking for the German dataset; this may be due to the fact that the amount of German fine-tuning data was relatively smaller than that of the French, and was therefore beneficially augmented by the French training data. In general, one can see the advantages of more fine-tuning data in two aspects: Firstly, the bigger French dataset provided more stable results; secondly, the sentence classification exhibits – also for German – a lot less variation, as it is only a two-class classification problem, which gives each class effectively more training data than for the NER task.

Concerning the hyperparameter maximum sequence length, the idea that a longer sequence could cover more context and thus perform better was not met at all. Generally, we found no clear pattern giving precedence to one maximum sequence length, especially for the entity recognition task; for sentence classification, the lengths 128 or 256 often got the highest results.

To sum up, although some trends are visible throughout the experiments, like the positive impact of more fine-tuning data or multilingual fine-tuning for multilingual models, no general lines emerge which would make it possible to give clear recommendations of which models perform well on which tasks. For example, the model bert-base-historic-multilingual trained in French and German featured very good results for the sentence classification task, but only played in the upper midfield for entity recognition. Still, there are a few models which provided satisfactory results for both tasks, which gives a good basis for the inference process on the *impresso* corpus. Before we came to this, we performed an error analysis to get an idea of the strengths and limits of the trained models.

[11]Token statistics per year for the French and German Europeana training corpora: `https://github.com/stefan-it/europeana-bert/blob/1.0.0/french_year_token_stats.png` (French) and `https://github.com/stefan-it/europeana-bert/blob/1.0.0/german_year_token_stats.png` (German).

4.2.2 Error Analysis

The error analysis aims to have a closer look at the performance of the models along different characteristics which might be important for the inference process on the *impresso* corpus. Knowing, for example, how well the model can correctly classify agency mentions with OCR noise helps to get an idea of the model's capacities to generalize to the noisier parts of the corpus.

We start with a plot of confusion matrices in the form of heatmaps for French (Figure 4.3) and German (Figure 4.4), which visualize models' performance per class (*bert-base-french-europeana-cased* for French and *bert-base-german-cased* for German). With maybe the exception of *DNB* in the French case, both models do not have many systematic false attributions to another agency. The classification of non-agency tokens as agency mentions is also rare; the most frequent error is the omission of true agency mentions. This is consistent with the results presented in the previous section, which generally showed higher precision than recall.

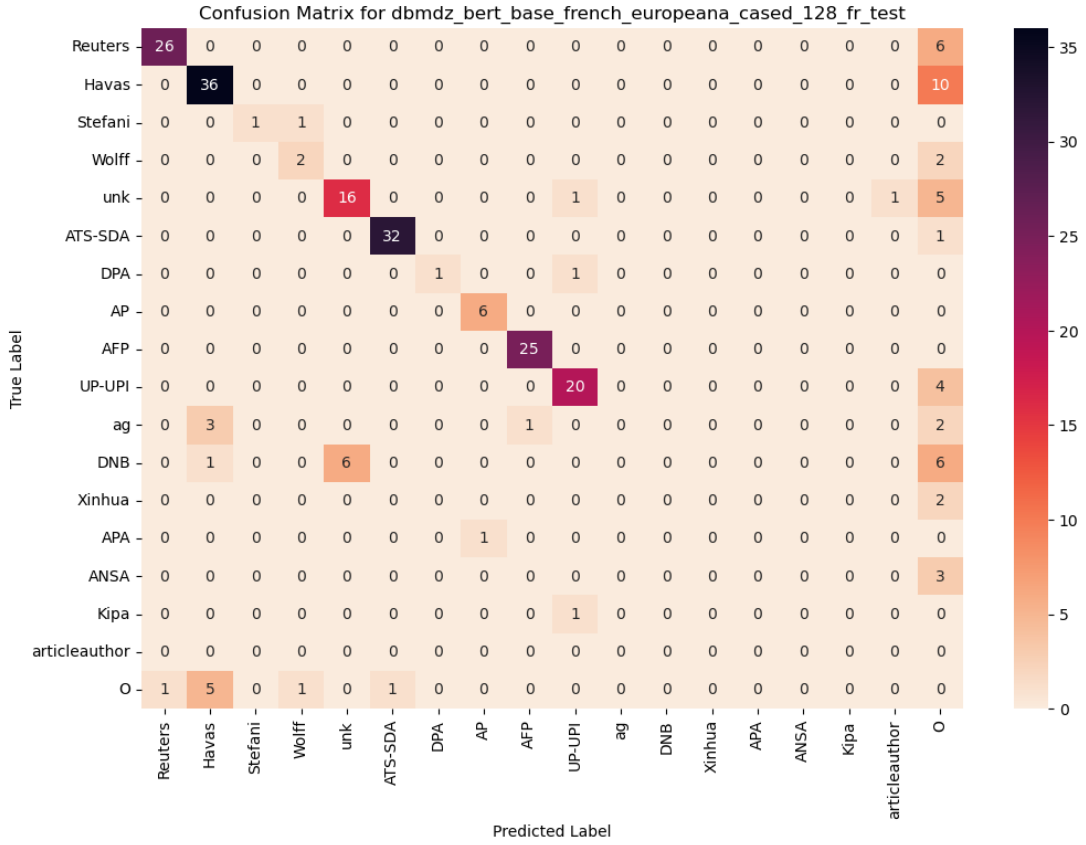


FIGURE 4.3
Heatmap showing the performance split by the different classes for the model *bert-base-french-europeana-cased* with a maximum sequence length of 128 on the French test set.

The HIPE scorer provides the functionality to split the results by time and noisiness of the classified mentions, which we used in the following analysis. Regarding OCR noise, we utilized the Levenshtein distance (LED, Levenshtein 1965) as a measure[12], which indicates no noise at LED0.0 and maximal noise at LED1.0. Figure 4.5 reveals that the F-score decreases dramatically if there is any noise in the mentions, both for French and German, although some models seem to cope slightly better with noisy mentions than

[12]The version of LED used here counts the number of alterations which are needed for one string to become its counterpart and then normalizes over the sequence length.

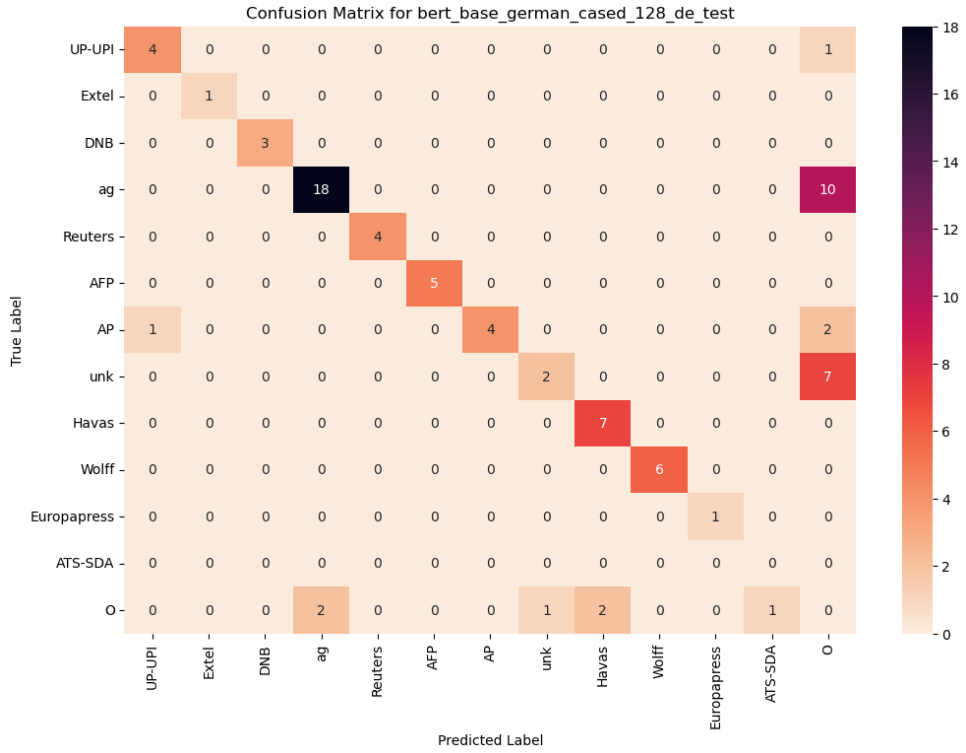


FIGURE 4.4
Heatmap showing the performance split by the different classes for the model *bert-base-german-cased* with a maximum sequence length of 128 on the German test set.

others. The level of noisiness – if the mentions feature noise at all – does not make a great difference for the performance, at least for French. As the German test set only contained one level of noisiness apart from zero noise, this analysis could not be reproduced for the German language.

Missing representative data over all subcategories also posed a problem for the splitting of performance scores per time. Both for the test set (see Figure 4.6) and the dev set (see Figure E.11), there is not enough data to make a sound assessment of the model's performance in the 19th century. While there is almost no data for German, the French test set hints at a worse performance of the classifier in the 19th century. However, the existing data in the dev set does not show this tendency, so the final evaluation remains unclear. On the other side, we found a clear trend regarding OCR noise, so if it is known that the OCR is worse for older newspapers, it can be deduced that the classification will be less strong. Over the 20th century, both languages show a relatively stable performance, which lets us conclude that in general, the models' capabilities do not necessarily increase over time.

For a selection of the best models, we also tested how well they correctly classify agency mentions based on the position of the mentions in the article. Figure 4.7 indicates that models are generally very good at finding news agencies at the beginning of the articles, and some models show equally good results at the end. So models seem to learn some kind of structure apart from the agency names.

Classification of Unseen Agencies

In order to understand to what extend the models rely on specific agency names and acronyms for their recognition, and how much they grasp sentence characteristics, we conducted a classification experiment with unseen agencies on a selection of models. For this, we swapped each agency with an agency unknown to the models, respecting the length of the agency mentions. Mentions with one token were substituted

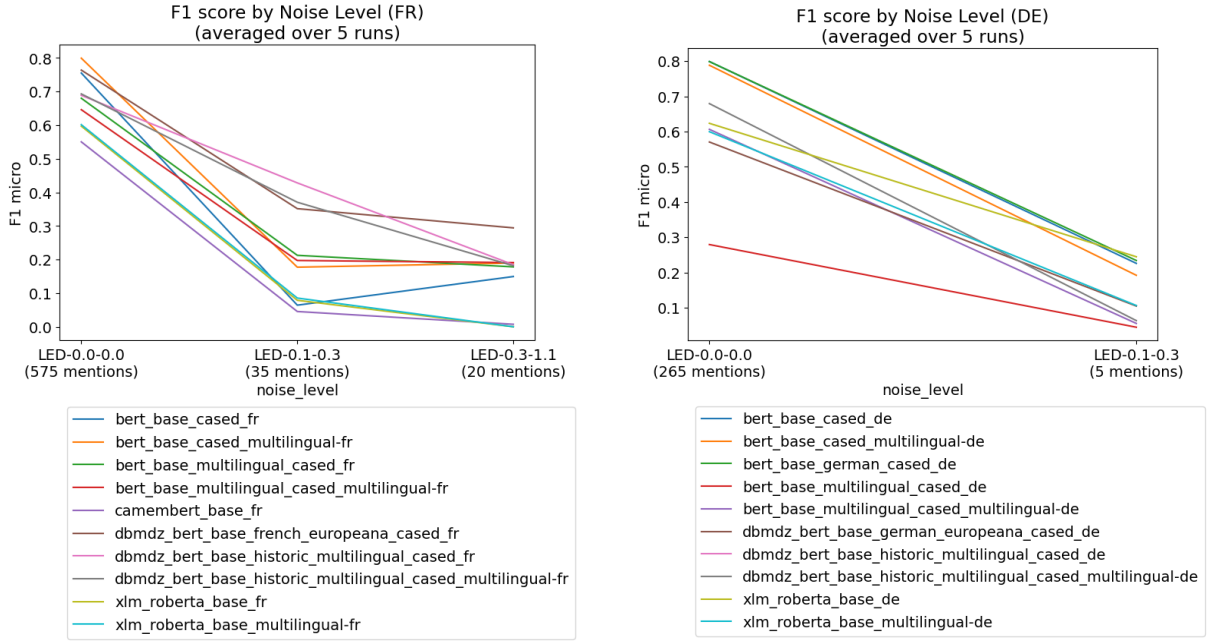


FIGURE 4.5
F-score of the different models on the French (left) and German (right) test set, split by the noise level of the agency mentions (measured with Levenshtein distance) and averaged over all configurations per model.

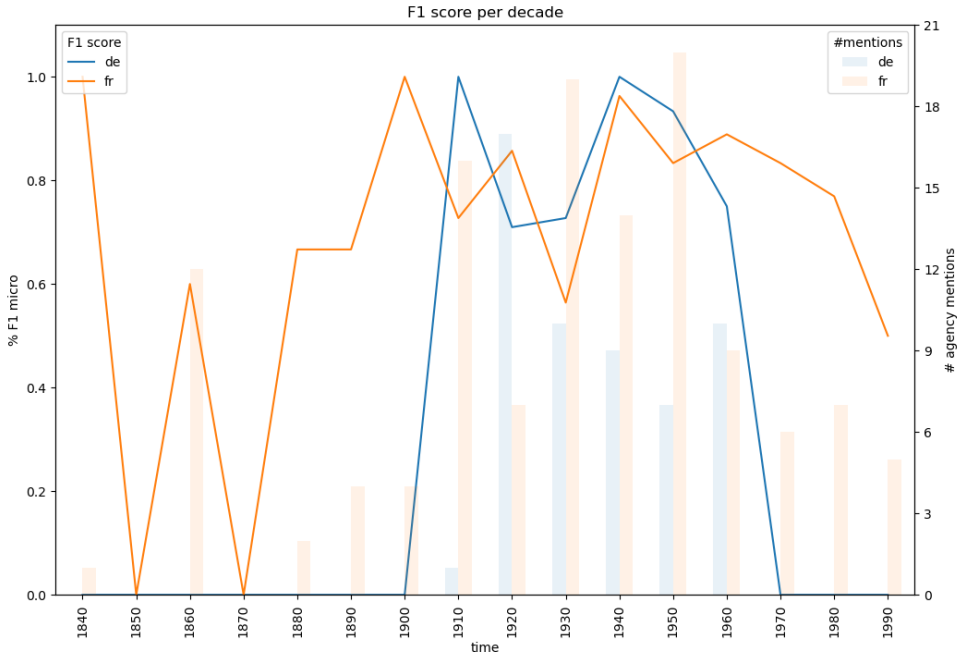


FIGURE 4.6
F-score of *bert-base-french-europeana-cased* (fr) and *bert-base-german-cased* (de) for the test set, split by decade. The bars in the back show the number of mentions existing in the respective test sets per decade.

with a new agency name consisting of one word, mentions with two tokens were swapped with two-word agencies and so on. As previously unknown agencies, we used the following:

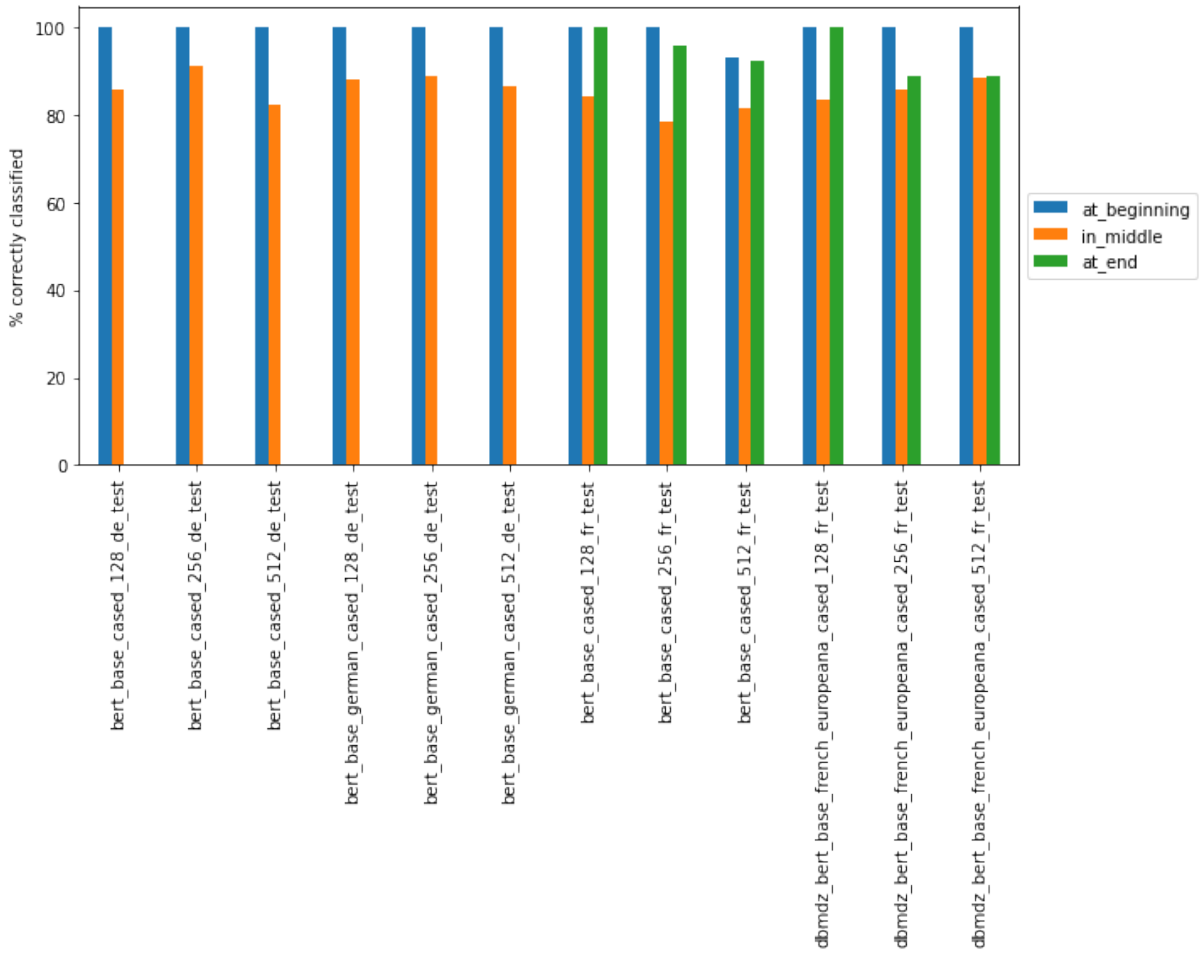


FIGURE 4.7
Percentage of correctly classified agency mentions on the test set per position in the article (at the beginning, middle and end), for a selected number of (well-performing) models. The German test set does not have agency mentions at the end, so numbers only exist for the beginning and the middle.

- One token: `PA`, `EPD`, `Fides`, `NTB`, `EFE`, `Fournier`, `AGERPRESS`, `SPT`, `ROSTA`, `CAPA`
- Two tokens: `Canadian Press`, `Samachar Bhavan`, `Agence Meurisse`, `Agence Balcanique`, `Agence Fabre`
- Three tokens: `Agence télégraphique ottomane`, `Athens News Agency`, `Agence de Constantinople`

Then, we let the models perform both the entity recognition and sentence classification task. For the former, it basically tested the classification capacity of the *unk* class, while the latter might provide more information on BERT's understanding of structural aspects of articles mentioning news agencies.

The results in Table 4.6 reveal a mixed picture. The sentence classification results are surprisingly high (maximal F-score of 0.976 for French and 0.980 for German), which indicates that the model indeed seems to learn from the structure and/or content of the phrases instead of only the finite set of agency names. In contrast, the performance for entity recognition is low, with maximal F-scores of 0.285 for French and 0.324 for German. As for the general results, the precision is higher than the recall. This is also confirmed by the heatmaps for class distribution (Figures 4.8 and E.12), which exhibit a very high

number of missed agency mentions. In both French and German, originally unknown agencies are often mistakenly classified as instances of Havas. This might be due to the fact that Havas was the agency with the noisiest mentions in both training sets (see Figure E.13).

TABLE 4.6
Performance for selection of models on test sets with swapped agencies unknown to the models; the maximum sequence length is 128.

Language	Model	Named Entity Recognition			Sent. Classif.
		F1	**Precision**	**Recall**	**F1**
fr	bert_base_cased	0.192 (0.042)	0.293 (0.045)	0.144 (0.037)	0.968 (0.002)
	dbmdz_bert_base_ french_europeana_cased	0.118 (0.063)	0.150 (0.081)	0.097 (0.052)	**0.976 (0.001)**
multilingual -fr	bert_base_cased	**0.285 (0.045)**	**0.382 (0.076)**	**0.227 (0.033)**	0.973 (0.003)
	bert_base_multilingual_cased	0.144 (0.089)	0.204 (0.125)	0.112 (0.07)	0.971 (0.003)
de	bert_base_cased	0.192 (0.039)	0.245 (0.064)	0.161 (0.032)	0.975 (0.004)
	bert_base_german_ cased	0.205 (0.093)	0.240 (0.125)	0.181 (0.074)	**0.980 (0.002)**
multilingual -de	bert_base_cased	**0.324 (0.086)**	**0.399 (0.115)**	**0.275 (0.073)**	0.971 (0.004)
	bert_base_multilingual_cased	0.190 (0.124)	0.232 (0.147)	0.162 (0.109)	0.969 (0.004)

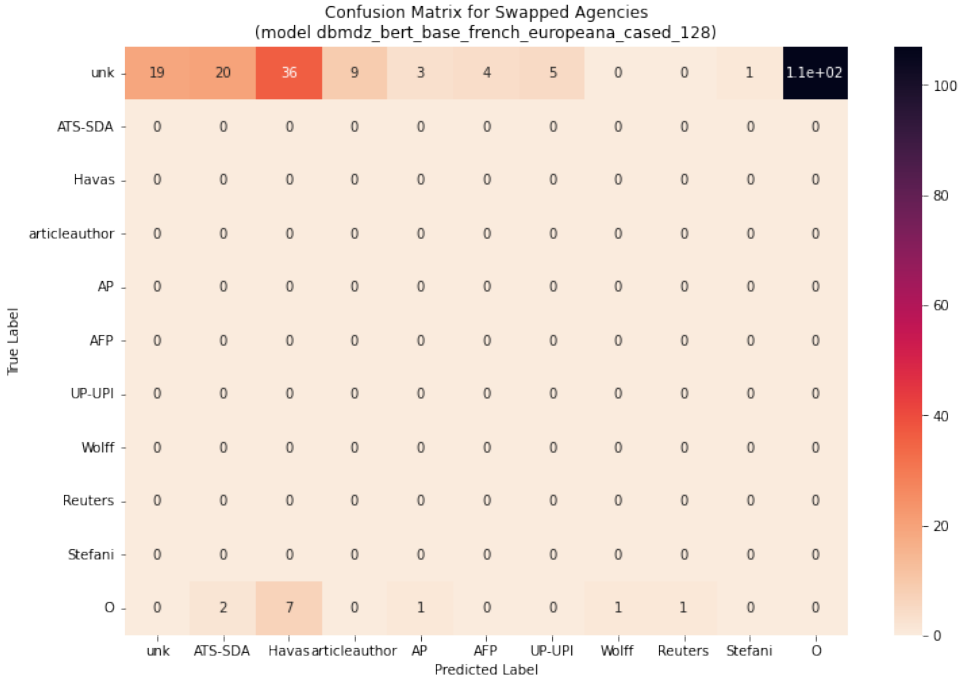


FIGURE 4.8
Heatmap showing the distribution over the different classes for the model *bert-base-french-europeana-cased* with a maximum sequence length of 128. Results concern the performance on the French test set, where the agencies were swapped with agency names unknown to the model.

4.3 Conclusions and Limitations

In this chapter, we conducted a number of experiments to train an optimal classifier for the agency detection task. We tried different model configurations and varied the hyperparameter maximum sequence length, which lead to diverse and sometimes surprising results (see 4.2.1). In general, the outcome was

satisfying, especially compared to the F-scores usually achieved for historical NER. Further improvements could have been attempted through data augmentation for the sentence classification, using the text reuse clusters from *impresso*, or by including the *impresso* word embeddings. A more thorough hyperparameter search might also have increased the scores, e.g. for models such as XLM-roBERTa.

One main limitation of the experiments was the lower resource of German fine-tuning data, which caused unstable results for many models. Moreover, the analysis of structural or linguistic characteristics was considerably limited by the focus on sentence-level instead of article-level classification. The way the model was trained, only sentences with an agency mention were considered as text with agency content, and not *all* sentences which were part of an article with an agency mention. Although BERT would not have allowed classifying an article as a whole, this problem could have been circumvented by labelling all sentences which were part of an agency article as "positive" in a pre-processing step. On the other side, the OLR for article boundaries is not always precise, which would lead to non-agency articles labelled as containing agency content. A classification based on these might produce unreliable results.

The error analysis in Section 4.2.2 confirmed a finding from the general discussion of model performance, namely that the models often either classify agency mentions correctly or do not detect the mentions at all, resulting in a higher precision and lower recall. We furthermore can expect models to work reasonably well throughout time (at least in the 20th century) and to safely detect agency mentions at the beginning and sometimes the end of an article. Regarding OCR noise, however, the models' performances decrease drastically if they encounter any noise in the mentions. Indeed, this problem is typical for historical NER (Ehrmann, Romanello, Flückiger et al. 2020, Ehrmann, Hamdi et al. 2023). Although the models seem to grasp structural aspects of agency content (see the experiments on unseen agencies 4.2.2), their entity recognition performance on the *unk* class is very weak. Thus, we expect the models to miss a lot of previously unknown agencies in an unseen text corpus. The training on more *unk* tokens might have increased performance for this subtask.

Although there exists a wide spectrum of possibilities to improve the models, some of them still seem to be solid classifiers which can be used for inference on the *impresso* corpus. We decided to select two models for this process, one for French and one for German. Regarding the hyperparameter maximum sequence length, we considered a length of 64 to be too low, as we would cut a lot of context because many sentences are longer than 64 tokens (see Section 4.1.2). The sequence length of 512 did not yield good results, so the decision lay between lengths 128 and 256, which performed comparatively well. As more computation is required for longer sequences (Devlin et al. 2019), we chose the maximum sequence length of 128. The final choice for the models fell on *bert-base-french-europeana-cased* for French and *bert-base-german-cased* for German, because they have high evaluation scores for both entity recognition and sentence classification, and promise to be stable configurations which might generalize well, as the different runs did not exhibit a lot of variation. Additionally, they are among the most robust models for mentions with OCR noise (Figure 4.5) and find all agency mentions at the beginning and the end of the articles (Figure 4.7).

Having chosen those two models, we proceeded with the next step, the inference on the *impresso* corpus.

5 News Agencies in the *impresso* Corpus

After having trained the deep learning models on the training data, we could finally run the two chosen models (see last chapter) on the whole *impresso* corpus. This chapter goes into detail on the technical side of the inference process and makes a quality assessment of the outcome, before concentrating on the analysis of the detected news agency articles. The analysis consists of a high-level overview of news agencies in the Swiss and Luxembourgish media ecosystem and a case study which examines the impact of German occupation on the news world in Luxembourg during the Second World War.

5.1 Inference on the *impresso* Corpus

The technical implementation of the inference process, i.e. the application of the two trained models (German BERT & French Europeana BERT, see Section 4.3) on all articles in the *impresso* corpus, was entirely done by Emanuela Boros and Maud Ehrmann. We included its specifications in this report for reasons of completeness and comprehensibility.

The inference was run on 24,994,906 French and German articles from *impresso*, using either German BERT or French Europeana BERT for the agency entity recognition task, depending on the language of the article at hand. In total, 4,482,890 agency mentions in 2,406,634 articles were detected. Table 5.1 provides statistics on the mentions split by country and language, while Figure 5.1 gives an overview of the number of articles with detected agency mentions over time, compared to all articles in the *impresso* corpus.

TABLE 5.1
Overview of the results of the inference process.

	Lg.	Agency Mentions	Articles with Mentions	Articles in *impresso* corpus
CH	de	1,488,570	549,614	4,513,041
	fr	2,838,159	1,741,604	17,759,345
Total		4,326,729	2,291,218	22,272,386
LUX	de	140,593	102,874	1,717,854
	fr	15,568	12,542	1,004,666
Total		156,161	115,416	2,722,520
All	de	1,629,163	652,488	6,230,895
	fr	2,853,727	1,754,146	18,764,011
Total		4,482,890	2,406,634	24,994,906

In the following, we first give technical details on the inference process and then continue with a quality assessment of its results, to give an idea about the reliability of the detected agency mentions.

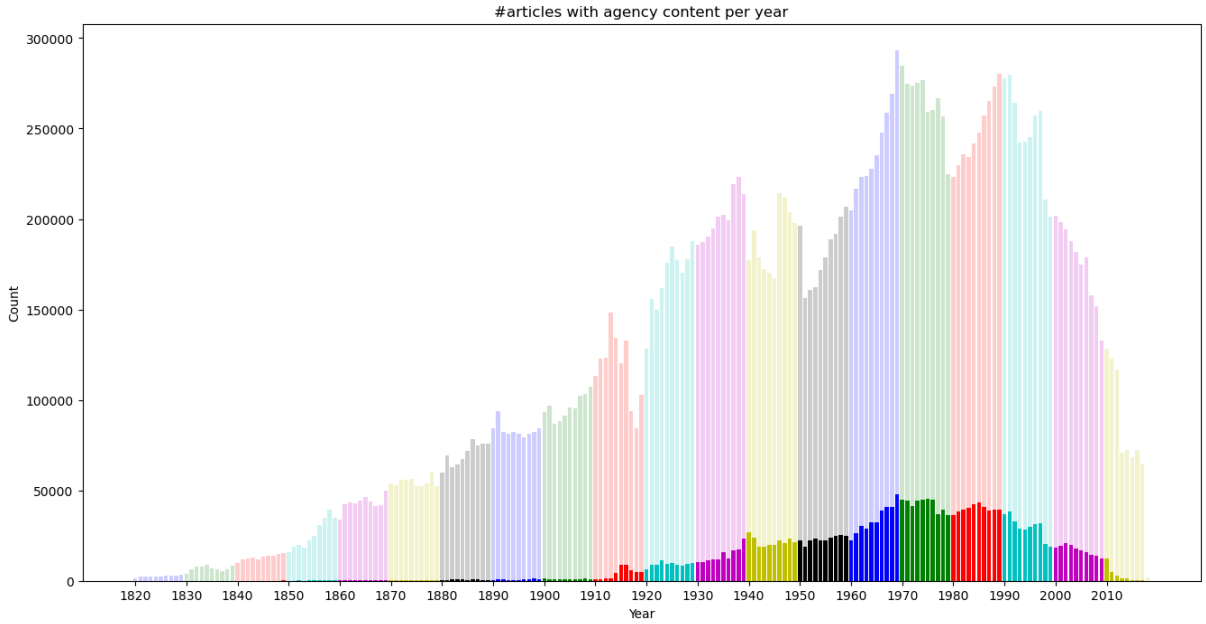


FIGURE 5.1
Number of articles with a detected news agency, compared against all articles in *impresso* over time.

5.1.1 Technical Details

In order to detect news agencies at a large scale, we took advantage of a framework built for processing documents in *impresso*, relying on two Python-based libraries, Dask[1] and TorchServe[2] for scaling machine and deep learning workloads[3].

Dask is a library for parallel computing in Python which enables the construction of complex, out-of-core computation workflows that maximize the computational capability of a Python ecosystem. It operates by building a task graph of the required computations, and then executes the graph in parallel, making use of all available computational resources, whether those are the multiple cores in a single laptop or a cluster of servers. This allows for the processing of larger-than-memory datasets or compute-intensive tasks.

In this project, Dask was complemented by TorchServe, which is specialized in PyTorch models, simplifying their deployment at scale in production environments. TorchServe achieves this by making it easier to package prediction models for inference and to manage them in a robust and scalable manner. Specifically, TorchServe uses a RESTful API[4] for both inference and management calls.

The framework utilises this toolset with TorchServe which focuses on serving prediction models, and Dask which focuses on scaling and managing the data at a large scale.

As shown in Figure 5.2, we first split the large dataset into manageable chunks of ca. 16,000-19,000 news articles each (encoded as JSON lines), resulting in 2,197 archives, which were saved physically on the disk. Next, these archives were handled by Dask with batch processing (Dask *bags*). Specifically, Dask took the list of archives as input and broke it into smaller batches (*bags*) which were then processed simultaneously with a number of workers.

[1] `https://www.dask.org/`

[2] `https://pytorch.org/serve/`

[3] The text of this section is based on notes provided by Emanuela Boros.

[4] A REST API (also known as RESTful API) is an application programming interface (API or web API) that conforms to the constraints of REST architectural style and allows for interaction with RESTful web services. `https://en.wikipedia.org/wiki/Representational_state_transfer` `https://en.wikipedia.org/wiki/API`

Meanwhile, TorchServe loaded both NER models, for German and French, and also started a number of workers to handle the parallel serving of models, which then waited for processing requests. Each worker in TorchServe operated individually, either using the German or French agency detection model for inference.

Dask handled the archives contained in a Dask *bag* in parallel, sending each document within the archives to TorchServe as an API request. TorchServe harnessed its waiting workers for the inference and returned a list of recognized news agencies to Dask, which were finally saved.

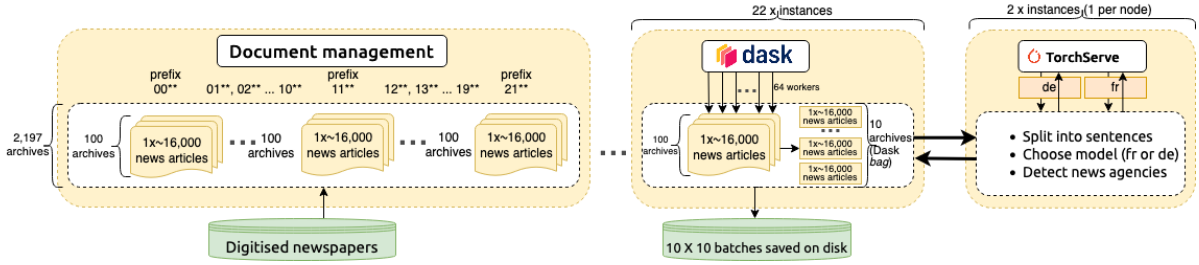


FIGURE 5.2
The pipeline of the process of inference at large scale for the *impresso* corpus.

Specifications. We used two cluster nodes with 8x A100 40G (NVIDIA Tesla A100) and 64 CPU cores each. The optimal number of workers, for both Dask and TorchServe, is 64 (as the number of cores). Further, we chose a batch size of 10 (the number of archives in a Dask *bag*). We ran several instances of Dask simultaneously, by further splitting the 2,197 archives into groups of 100, thus 22 processes sent NER requests to two TorchServe instances (one per cluster node). The process took two weeks.

5.1.2 Quality Assessment

To assess the quality of the results, we checked a randomly sampled subset of tokens classified as agency mention in the *impresso* corpus. The subset consisted of 160 tokens, five per decade (1840–1990) and language (fr, de).

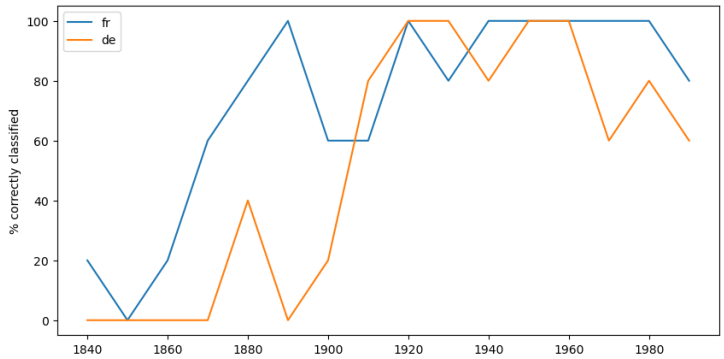


FIGURE 5.3
Percentage of correctly classified tokens in the checked sample, split by time and language.

Figure 5.3 displays the outcome of the assessment over time. It shows that for both languages, the classification in the first decades is not reliable, but the quality improves rather quickly for French, with all five verified samples being true positives in the decade 1890. In contrast, the German sample for the same decade features no true positives. However, after 1910, the classifiers for both languages show a strong performance, promising reliable results for the classification as a whole. Only the German curve drops

slightly from the 1970s on, which coincides with the number of training samples the German classifier has seen during training for this time (see Figure 3.12).

A closer look at the wrongly classified tokens in the subset (Figure 5.4) reveals that many of those tokens were no news agencies at all. The labels given to the tokens are relatively diverse, although *unk* and *Havas* appear most frequently.

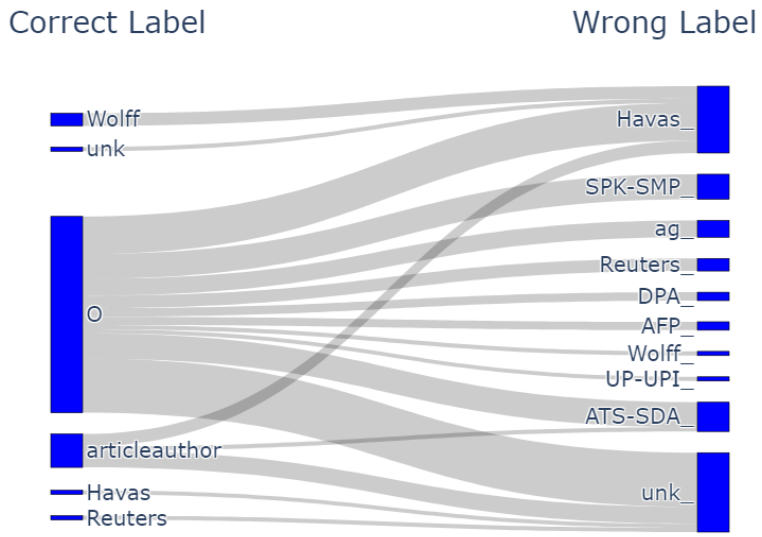


FIGURE 5.4
Predicted and correct class for the wrongly classified tokens in the checked sample.

Concerning *Havas*, most wrongly classified tokens occurred before 1880 (see Figure F.1 in the Appendix), which corresponds to the quality of the classifications in general. It is interesting to note that several *Wolff* mentions were classified as *Havas*, all part of the French corpus. This might stem from the low number of *Wolff* labels in the French training corpus.

In general, the wrong classifications show certain structural patterns: 52% have brackets around them, 64% are followed by a dot and 7% have a hyphen in the vicinity. Only 23% exhibit neither of the three characteristics, which indicates that they played a role in the classification process.

The wrongly classified *unk* token motivated us to inspect its distribution on the whole corpus further. The top ten tokens classified as *unk* are (with their frequency in brackets):

. P. (12,656), Fournier, (9,027), Bureau (7,708), FN (7,316), D. N. B. (7,163), . T. S (3,478), Telunion (3,332), Korrespondenz (3,053), . N. B. (2,981) and Cosmo (2,927).

While the tokens Fournier and Telunion belong to agencies and thus were (most probably) classified correctly as *unk*, the remaining classifications are questionable or wrong. The token D. N. B. should have gotten the label *DNB*, and . N. B. and . T. S are noisy mentions for *DNB* and *ATS* respectively. A closer examination of . P. uncovers that it is mostly part of C. P., i.e. *Correspondance Particulière*, but sometimes also of D. P. A., so it partially contains agency mentions. The same holds for Bureau and Korrespondenz, which sometimes appear in agency names, but are generic enough to be interspersed with wrong classifications. The token FN only appears in the newspaper "Freiburger Nachrichten" and thus is an acronym for their paper, not an agency. All in all, the *unk* class provides a mixed picture, with some correct classifications, but also many false positives.

Finally, we looked at the most common tokens which were predicted to be agency mentions. Figure 5.5 lists the top 40, suggesting that the majority of the classifications are correct, as most tokens are names of agencies.

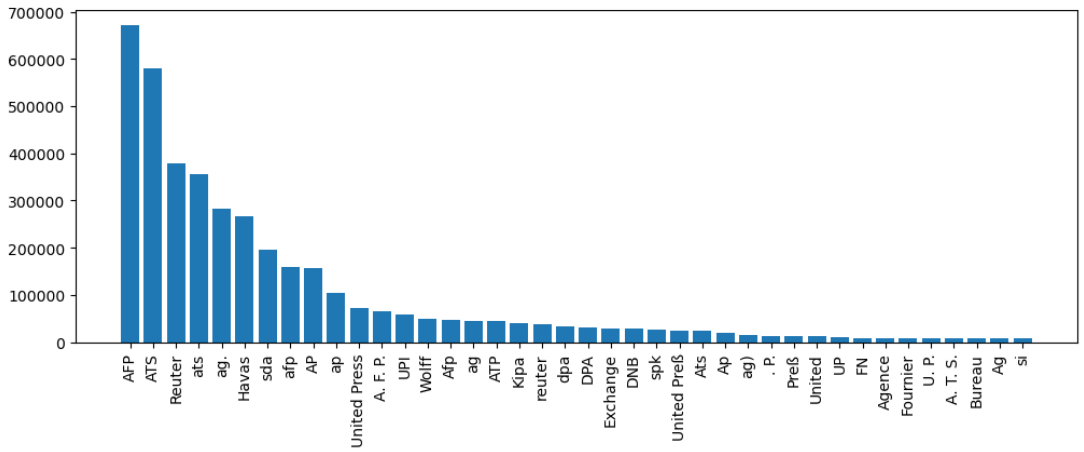


FIGURE 5.5
40 most common tokens predicted to be an agency mention in the *impresso* corpus.

Regarding the exceptions, `.` `P.` and `FN` were already discussed above. A check of some samples of `ATP`, which were usually classified as *AFP*, reveals that they mostly belong to the *Association of Tennis Professionals*. This finding is supported by the fact that most tokens appear after the association's foundation in 1972 (see Figure F.2 in the Appendix). Although the token `si` did not appear before in the classification process, it belongs to the Swiss sports news agency *Schweizer Sportnachrichtenagentur Sportinformation*, which was founded in 1922. Unfortunately, it is mostly classified as *ATS-SDA* instead of the correct *unk* label.

To conclude, the quality assessment mostly mirrors the findings from the error analysis in Section 4.2.2. The overall classification provides good results, with severe quality issues in the 19th century, but a strong performance in the 20th century – although some false positives always need to be anticipated. In contrast, the results of the *unk* class are unreliable and need to be treated with care. However, the class contains true agency mentions, so it could be used for research if it was complemented by a search query for a specific agency such as *Fournier*.

5.2 News Agencies in the Media Ecosystem

After the successful application of the agency detection on all articles in the *impresso* corpus, we could analyse the outcome and finally get an idea of the role news agencies played in the Swiss and Luxembourgish media ecosystem. How much content can be traced back to news agencies? When did newspapers begin to systematically cite news agencies? Which agencies were most influential, and which newspapers relied on which agencies? These are some of the questions which will be addressed in the following section. For this, we first rely mainly on descriptive methods in Section 5.2.1 and then continue with a network analysis in Section 5.2.2.

5.2.1 News Agency Content in Swiss and Luxembourgish Newspapers

Regarding the presence of agency content in the *impresso* corpus, Figure 5.1 already gives a first impression of the development over time. Table 5.1 shows that most agencies were detected in Swiss French newspapers, which is due to their high amount of articles in the *impresso* corpus. For the Luxembourgish

part, although both languages are almost equally represented in the corpus, the amount of detected agency content in German newspapers is significantly higher. Based on the assumption that the quality of the classification is good in the 20th century, we can conclude that French-speaking newspapers did not credit agencies as systematically as their German-speaking counterparts. However, the statistics were certainly also influenced by the fact that most newspapers in the Luxembourgish corpus are only available until 1950 and that there were almost no articles published in French during the German occupation in 1940-1944. We further address this aspect of the case study in section 5.3.

Newspapers' Reliance on Agency Releases. To get an idea of how much newspapers relied on agency releases for their articles, we computed the share of agency content in the *impresso* corpus over time. Figure 5.6 shows the results: Keeping in mind that the 19th-century agency detection is not very reliable, it still seems that newspapers did not regularly credit agencies at the beginning, systematic citation of agencies appearing only in the 1910ths. The first significant increase happened in the years 1914-15 and the second in the years 1939-40, indicating that newspapers relied heavily on news agencies for war reporting (M. B. Palmer 2019, p. 79ff). An analysis of the covered topics during wartime could shed further light on this circumstance, but this is out of the scope of this study. Another hypothesis would be that the disclosure of the source of information became more important, since reports from the war were heavily influenced by propaganda, also – especially – by news agencies (M. B. Palmer 2019, p. 76). This hypothesis could be investigated e.g. by comparing the share of cited agency content with the share of non-cited agency content, which could be approximated with the help of the *impresso* text reuse clusters.

The share of detected agency content reaches its peak in the 1970s, before decreasing by almost 10 percentage points between 1985 and 2000. This decline is surprising, since studies on the current role of news agencies in the media landscape suggest a high usage of agency content today (see Section 2.3). Possible explanations could be that the *impresso* corpus is less representative due to a lower number of available newspapers during this time, or that there was a change in the way agencies were used or credited. Less variability of agency content could also play a role, as only ATS-SDA and AP provide content on a national level in Switzerland since 1994 Meier 2010.

Still, since 1940, the percentage of articles with at least one detected news agency roughly lies between 10 and 18%. Taking into account that the precision of the agency detection was around 0.8 during the experiments, suggesting that around 80% of all articles with an agency mention have been found, we can estimate that around 13 to 22% of all articles in the corpus contain a reference to a news agency. Of course, these are only preliminary results on the proportion of agency content in the corpus, as the information on the text reuse clusters in *impresso* has not been incorporated yet – it will be interesting to see how much the numbers will increase when all the articles which have a textual overlap with a detected agency article are considered as well. Compared to the numbers presented in the literature review in Section 2.3, the 13-22% of explicitly credited agency articles of the *impresso* corpus lie between the 1% in the British media in 2016 (Lewis, Williams and Franklin 2008) and the 33.5% in seven Swiss newspapers, which have been calculated in the study by Vogler, Udris and Eisenegger (2020) for the years 2012-2018. However, due to the reliance on explicit references to news agencies, these numbers are very conservative estimates (e.g. while only 1% of the articles contained a reference to an agency in the corpus explored by Lewis, Williams and Franklin (2008), they could attribute almost half of the inspected articles to an agency release). Most studies rely on text reuse to also retrieve the agency content which was not explicitly referenced, usually finding that around 30-50% of the articles come from a news agency. After the inclusion of the text reuse information from *impresso*, it will be interesting to see where the ratio in this corpus lies, since it will be a first estimation of the reliance of newspapers on news agency content in the past.

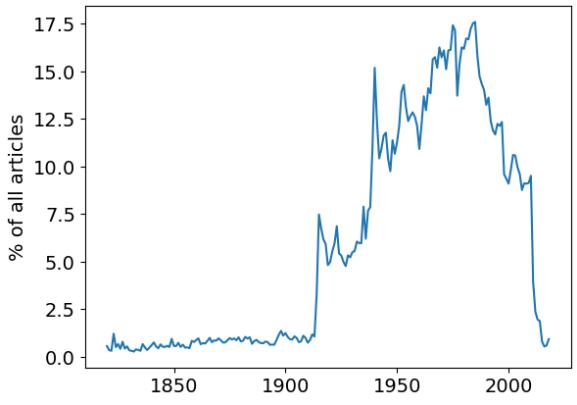


FIGURE 5.6
Percentage of articles in *impresso* where a news agency was detected, development over time.

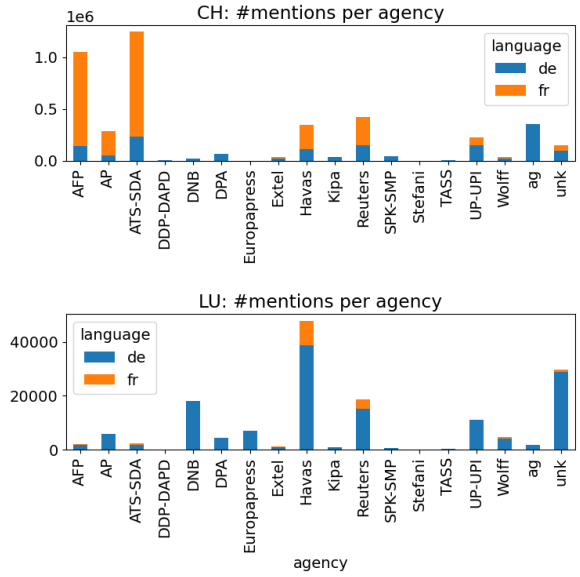


FIGURE 5.7
Number of agency mentions per agency, split by country (CH above, LU below) and language (German in blue, French in orange).

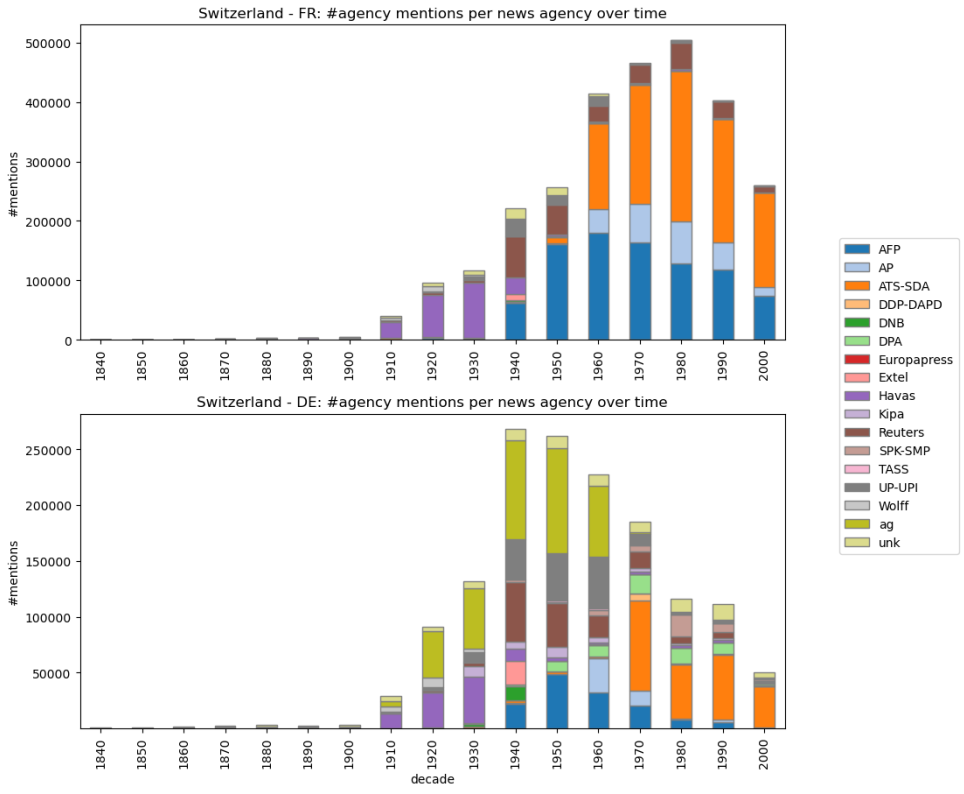


FIGURE 5.8
News agency mentions in Swiss newspapers over time, split by language (French above, German below).

News Agencies across Countries and Language Borders. Figure 5.7 presents the use of agency content across countries and languages. For Switzerland, the French AFP and Swiss ATS-SDA have

the greatest share of agency mentions, followed by Havas, Reuters and *ag*. Except for *ag*, the agencies mainly appear in the French part of the Swiss corpus, which can be explained by the high share of French newspaper articles in the corpus. The picture looks completely different for Luxembourg, where Havas dominates the corpus, followed by *unk*, the Nazi agency DNB and Reuters. However, these numbers cannot be compared to the Swiss distribution, as the Luxembourgish articles mainly date from before 1950, and most agency mentions occur between 1910 and 1950 (see Figure F.3) – thus the high share of Havas and the DNB, which only existed in the years 1933-1945. Still, the high proportion of the French Havas in German-speaking newspapers raises the question of why the agency was not cited in French-speaking newspapers as well, since it clearly was present in the Luxembourgish market. An explanation could be that the French newspapers did not systematically credit news agencies (at least until 1950), an idea which could be assessed with the help of the *impresso* text reuse clusters.

A closer look at the distribution of agency mentions in the Swiss corpus in Figure 5.8, split by language and time, reveals great differences. While both German- and French-speaking newspapers resorted to news from Havas at the beginning, only the French-speaking newspapers stayed with its successor AFP. Regarding the American news agencies, AP makes up a respectable share of mentions in the French-speaking part, but is only one of many news agencies in the German corpus, whilst UP-UPI is more present there. In general, the agency mentions in the German-speaking part are a lot more diverse, giving presence to agencies which rarely appear or are completely missing in the French corpus. In part, this might also be due to a deficiency of the French classifier, which did not see relevant agency names during training time. For example, the one instance of `P.S.M.`, the French version of SPK-SMP, was wrongly annotated as *unk* in the training set. Nevertheless, the observation that the French-speaking Swiss newspapers cited a less varied spectrum of agencies should still hold, since the French classifier saw most of the news agency classes during training time (for the distribution in the training set see Figure 3.13), and the *unk* class is not very frequent in the detected agencies during inference (Figure 5.7 and 5.8).

In both languages, the Swiss agency ATS-SDA is largely cited in the latter decades, although it appears considerably later in the German-speaking part of the Swiss corpus. Instead, the generic *ag* has the highest share of agency mentions in the German-speaking newspapers for a long time, and only decreases with the rise of ATS-SDA. A more detailed investigation of the *ag* class reveals that it was mainly used by three newspapers, the *Freiburger Nachrichten (FZG)* (75%), *Die Tat (DTT)* (19%) and the *Neue Züricher Zeitung (NZZ)* (5%). Plotting the distribution of news agencies in the two newspapers FZG and DTT over time in Figure 5.9, it is very striking that *ag* is heavily used, but then disappears within a few years, at the same time as ATS-SDA begins to be cited. This development can be discerned for both newspapers, albeit a bit less pronounced for DTT. It stands to reason that those newspapers used the generic *ag* as a descriptor for the agency ATS-SDA and only began to cite the agency under its name after 1970. We sampled ten articles which have been classified as using *ag* to test this hypothesis. Indeed, five out of them contain Swiss news, while only one article provides foreign (American) news. Two articles are false positives, and the other two use the abbreviation together with another agency descriptor, i.e. `ag. (DPA)`. So our guess is that we can attribute two uses to the `ag.` token, the placeholder for ATS-SDA and, although less often, the abbreviation of a generic "Agentur".

Concerning the presence of agency content in contemporary Swiss newspapers, Vogler, Udris and Eisenegger (2020) found that "the study at hand shows that editorial coverage is much more unique than agency-based coverage. [...] Our analysis showed that in Switzerland, the concentration of agency-based content is especially high for news on the national level." This suggests that today, newspapers rely heavily on ATS-SDA to produce their stories. Figure 5.8 already hints at a similar tendency in the second half of the 20th century, but we would need a closer content analysis, maybe with the help of the topic modelling from *impresso*, to make sound statements on this.

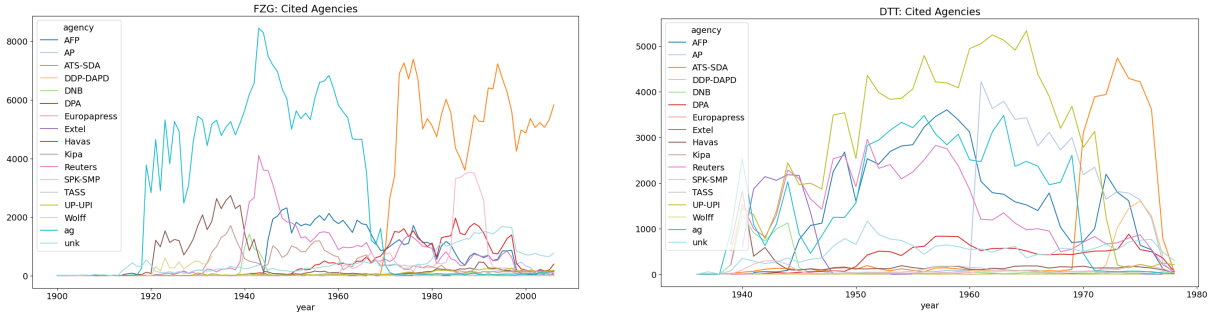


FIGURE 5.9

The presence of news agencies in the Swiss newspapers *Freiburger Nachrichten* (left) and *Die Tat* (right) over time. Those two newspapers make up 95% of all mentions of *ag* in the *impresso* corpus.

Agency Lifecycles. Figure 5.10 presents the frequency of nine news agencies in the *impresso* corpus over time. Generally, we observe that except for DPA, no agency was systematically credited before it existed, which speaks to the quality of the agency detection models. On the other hand, many agencies were only credited long after their foundation, as was the case with ATS-SDA, Havas, Wolff and Reuters. This certainly is partly due to the fact that agencies were seldom systematically cited in the 19th century; the moderate quality of the classifiers during this time might also play a role. However, in the 20th century, the results are interpretable and show some interesting characteristics.

Comparing the curves of ATS-SDA and *ag*, we again see that ATS-SDA emerges when *ag* decreases, following the reasoning from the paragraph above. The French agencies Havas and AFP exhibit a similar development, although even more abrupt, which suggests that AFP took over the position of Havas in the market in 1944. The network analysis in the next section will second this. In the years 1940-1944, Havas did not exist officially any more, but was turned into the state agency *Office français d'information (OFI)* by the Vichy government. Still, Havas was cited in the *impresso* corpus, as its graph in Figure 5.10 reveals. A check of some articles shows that indeed, many newspapers continued to cite OFI as Havas, or turned the agency reference into `Havas-OFI`.

Regarding the British agency Reuters, sources indicate that the agency insisted from the beginning that they should be credited in the newspaper articles (Boyd-Barrett and M. Palmer 1981, p. 39f), even giving discounts if the newspapers followed their wishes (M. B. Palmer 2019, p. 17). However, we cannot see this reflected in the *impresso* corpus, possibly because Reuters was not present in Switzerland during the times of the cartel (and most of the articles in the corpus are from Swiss newspapers). According to Shrivastava (2007, p. 5), Havas and Wolff divided the Swiss market between them until the foundation of ATS-SDA in 1894. The *impresso* corpus gives evidence of this, as the American agency AP, which participated in the cartel as well, only appeared slightly around 1950 and really established itself after 1960. UP-UPI's challenge of the dominance of the cartel, described in Section2.1, can equally be retraced in Figure 5.10, because it rises to prominence much earlier in the corpus than Reuters and AP. The company's decline in the 1970s is mirrored in the corpus as well.

A comparison to the curve of AP raises the question of whether UP-UPI lost its Swiss and Luxembourgish clients to its biggest American competitor. Figure 5.11 disconfirms this guess. Solely the *Gazette de Lausanne (GDL)* seems to have changed its subscription from UPI to AP, while *Die Tat (DTT)* made use of both. Otherwise, the two agencies had different clientele: UP-UPI does rarely appear in AP's main customers *l'Express (EXP)* and *l'Impartial (IMP)*, and AP seems to have acquired *La Liberté (LLE)* one decade after the newspaper stopped citing UP-UPI.

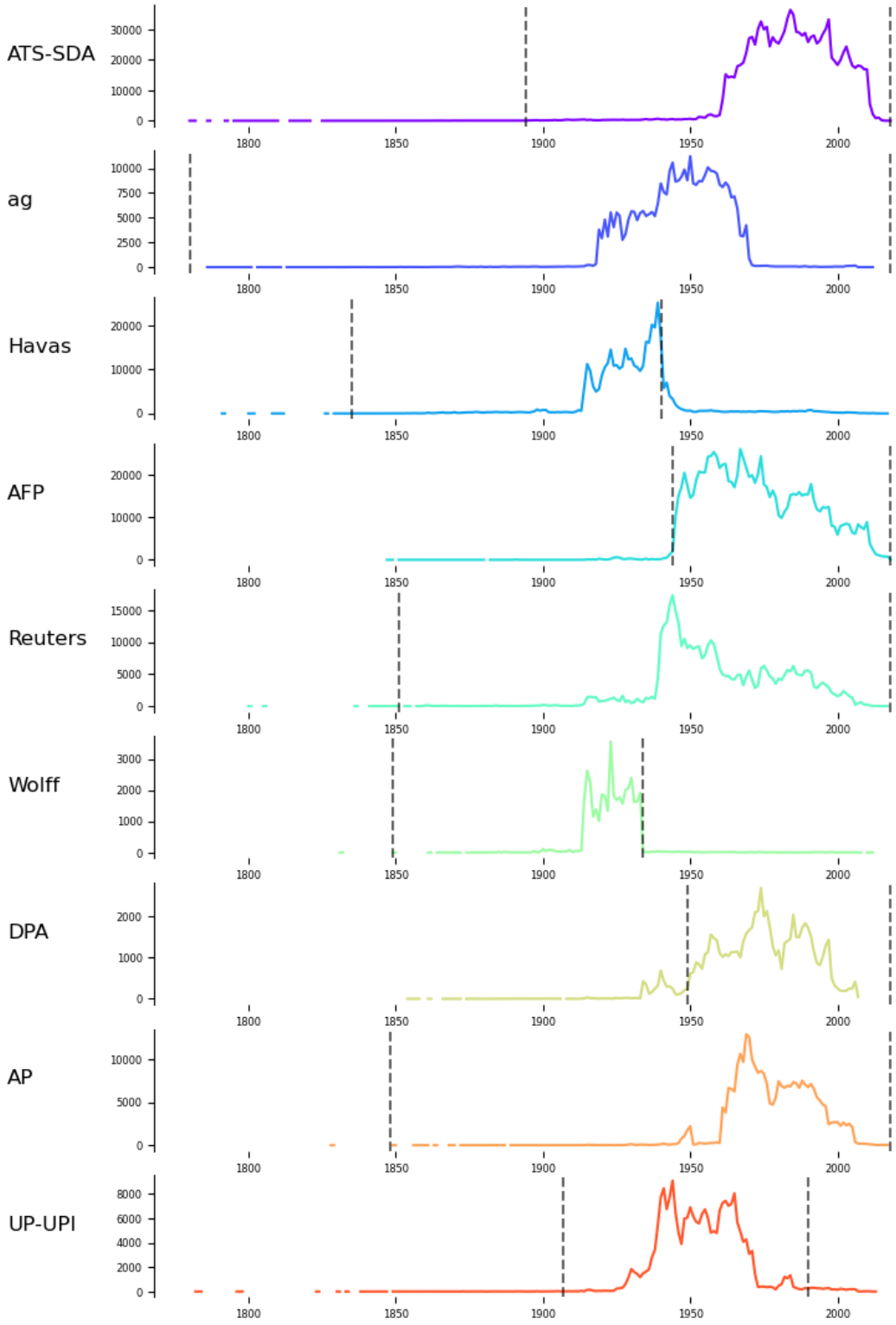


FIGURE 5.10
Number of mentions per news agency over time, for a selection of agencies (graphs for the other agencies can be found in the Appendix F.2). The absence of a coloured line means that no agency mentions were detected for the respective year. Each dotted line on the left marks the official founding year of the agency, and on the right dotted line its liquidation. For agencies still present today, the right line was set to 2018, the last year in the *impresso* corpus. Note that the y-axes differ with respect to each agency.

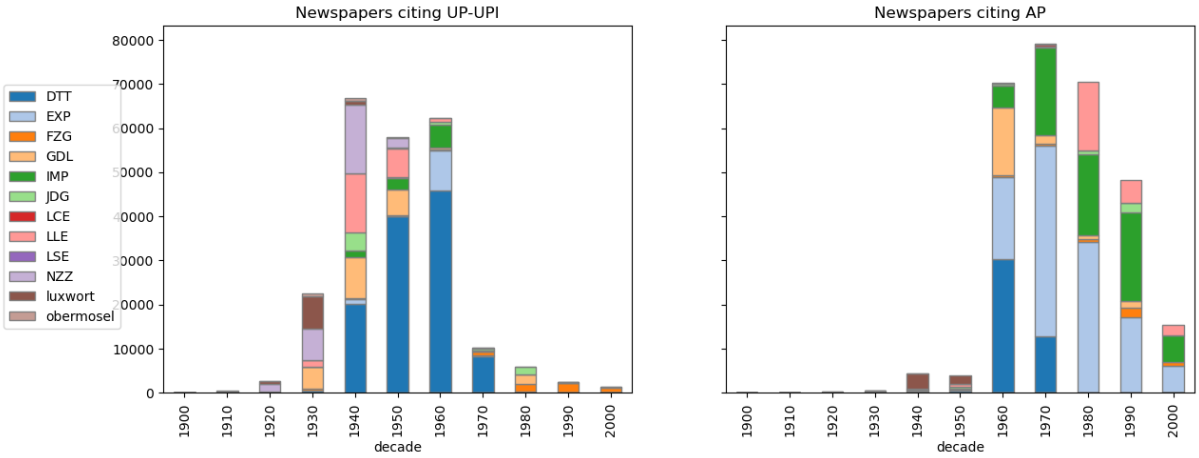


FIGURE 5.11
The presence of the American agencies UP-UPI (left) and AP (right) in Swiss and Luxembourgish newspapers over time. Newspapers with less than 100 mentions per decade were dropped.

Comparing Newspapers. Until now, we either investigated the general development of news agencies in the corpus or compared agencies to each other. Another option is to contrast newspapers' subscriptions to agencies. We will only give an impression of the research possibilities in this direction, by looking at the two French-speaking newspapers *Journal de Genève (JDG)* and *Gazette de Lausanne (GDL)*. Their plots of agency mentions across time in Figure 5.12 illustrate that the GDL started to credit agencies one decade earlier than the JDG – which also resulted in a higher share of references to Havas – but generally relied less on agency content in most of the decades (or explicitly marked agency content less). Apart from this, the distributions closely resemble each other, both newspapers mainly citing AFP, ATS-SDA and, to a lesser extent, Reuters. Two minor differences occurred in the 1940s when JDG relied on the British agency *Exchange Telegraph*, an agency with an original focus on financial topics but which offered war reporting (M. B. Palmer 2019, p. 63), and the 1960s when GDL referenced AP. Apart from the geographical proximity, the resemblance might be explained by the newspapers' similar political orientation (Clavien 2010). This presupposes that different agencies have different political tendencies, which is known for SMP-SPK (Windlinger 2011), but is not as obvious for other agencies, as they usually strove for impartiality (Boyd-Barrett and M. Palmer 1981, p. 36). Still, agencies could vary in the focus they lay on different topics and regions, e.g. if they operated regionally, nationally or internationally. A deeper historical study could shed light on these variations and how they translated to different uses in the newspapers. Additionally, returning to the comparison between GDL and JDG, the hypothesis that the two newspapers preferred similar news agency content could be tested by checking if the two newspapers published the same agency releases or if they picked up similar news stories, again using the methods of text reuse and topic modelling. The striking analogy of agency distributions in the 1980s warrants further investigation as well. In contrast, the similarity in the 1990s is no surprise, because GDL was bought by JDG in 1991, and the name "Gazette de Lausanne" only appeared as a subtitle to JDG until their merger with a third newspaper to become *Le Temps* in 1998.

5.2.2 The Network of Newspapers and News Agencies

This section follows yet another research path by examining the network of news agencies and newspapers, retracing the flow of information in the Swiss and Luxembourgish news world. As we did not incorporate text reuse clusters into the analysis so far, the network can only display the frequency of news flows from agencies to newspapers, instead of tracing news from newspaper to newspaper or following specific news through time, as was done by Salmi et al. (2020). The network based on the present data can still give

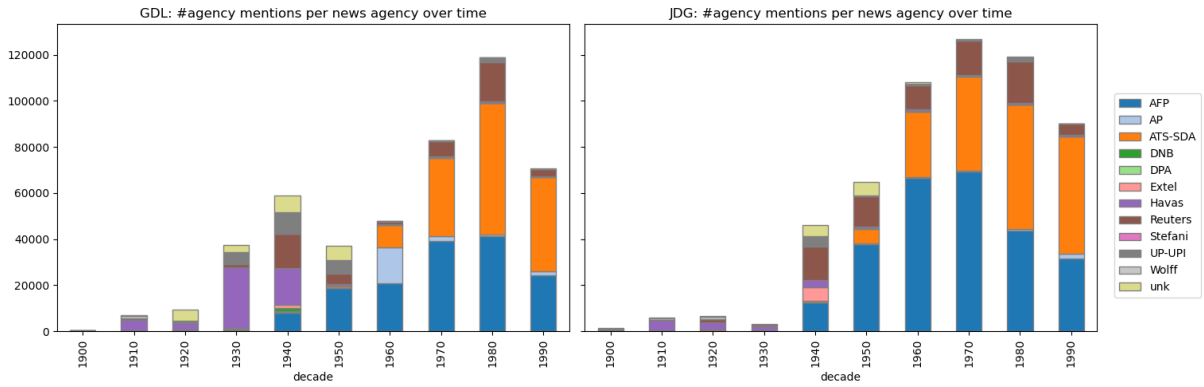


FIGURE 5.12
The presence of news agencies in the Swiss newspapers *Journal de Genève* (left) and *Gazette de Lausanne* (right) throughout the decades.

valuable information about the interconnectedness and structure of the news market over time.

In the following network, newspapers and news agencies form the nodes, while a link is created between them if a newspaper cites an agency. The result is a multigraph (it can feature several edges between two nodes), bipartite (newspapers and agencies form two separate node sets which have no links within their respective set) and directed (news only flow from agencies to newspapers). Technically, we will work with an undirected graph, as the directedness does not provide new information in the current state of the project, because the information only flows from agencies to newspapers. With the incorporation of text reuse data, this will change, as links between two newspapers or two agencies will become possible, making it important to distinguish the source and the recipient of the communicated information.

Figure 5.13 visualizes the network during six selected decades, outlining its development over time. To ensure comparability between the different views, the thickness of the links, which represents the quantity of shared news between two nodes, is normalized by the number of articles during the respective decade. The size of the nodes indicates the number of neighbours it has, i.e. how many other nodes it is connected to. In our context, a large newspaper node means that the newspaper cites many different agencies, while a large agency node signifies that the agency occurs in many different newspapers. We chose to start with the decade 1890-1900, as this is one of the first decades with reliable classifications. We can see a dominance of Havas, which appears in most of the newspapers, although only the *L'indépendance luxembourgeoise (indeplux)* cites it often. A look into some articles of *indeplux* shows that this newspaper already seems to credit agencies systematically. Wolff and Reuters are also present in the network, although the latter only is referenced by Luxembourgish newspapers. We can equally witness the beginnings of ATS-SDA, which was founded in 1894.

The network for the decade 1910-1920 already exhibits a lot more connections, which coincides with the increasing share of agency content in the *impresso* corpus. Apart from UP-UPI and the German agency Europapress, which enter the Luxembourgish market and also get cited by the Swiss-German *Neue Züricher Zeitung (NZZ)*, the news agency landscape stays roughly the same. Especially Havas is now heavily cited by the NZZ and counts most of the French-speaking Swiss newspapers among its frequent users of agency information.

Compared to other decades, the network for the years between 1935 and 1945 is extremely dense, revealing that newspapers relied on a great number of different agencies for their news during this time. Our hypothesis for this is that newspapers wanted to draw on as many sources as possible for their news during wartime, in order to get a comprehensive picture of the international situation, although this would need to be investigated further through deeper historical research. The citations include the

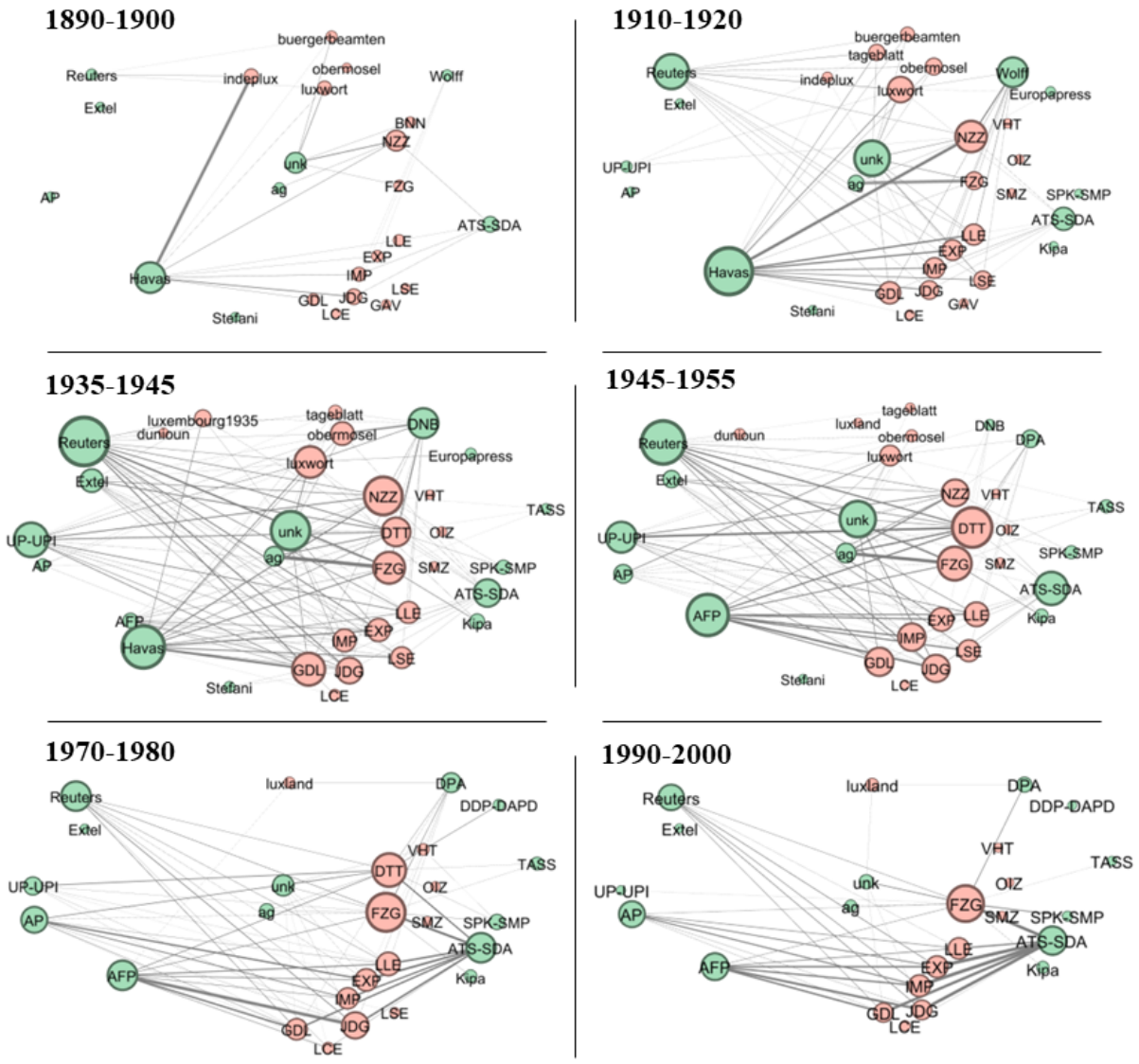


FIGURE 5.13

The network of news agencies (green) and newspapers (red) in six different time periods. An agency shares a link with a newspaper if it was referenced at least 24 times, the thickness of a link refers to the number of mentions (normalized by the total number of articles during the time). The size of a node is proportional to the number of links it has. The newspapers are clustered based on country and language: French papers from Luxembourg are on the top left, German Luxembourgish papers on the top right; German papers from Switzerland are on the right in the middle, while French Swiss papers can be found at the bottom. In general, Luxembourgish newspapers are written in lowercase, while their Swiss counterparts are displayed with uppercase letters.

Nazi-German agency *DNB*, which appears both in Luxembourg and Switzerland, though slightly less in the French-speaking newspapers. Moreover, the thick links between Reuters and many Swiss newspapers indicate that Reuters is now fully established in the Swiss news market. AFP, which was founded in 1944, emerges as well and rises to prominence in the years 1945-1955. A comparison of the links of AFP in the latter decade to those of Havas between 1935-1945 confirms the earlier hypothesis that AFP took over the place of Havas, as it features links to nearly all newspapers which had cited Havas before. With *L'Impartial (IMP)*, AFP even acquired a new client. In general, the Swiss newspapers seem to rely

on agencies between 1945 and 1955 as heavily as during the times of war, while the Luxembourgish newspapers *Escher Tageblatt (tageblatt)* and *Obermosel-Zeitung (obermosel)* return to a lower share of explicitly referenced agency content.

The network in the 1970s gives the impression to be less dense, which is partly due to the disappearance of most Luxembourgish newspapers and the NZZ, as well as the decline of UP-UPI and some other smaller agencies. However, newspapers also rely on less varied agency content, like *Die Tat (DTT)* or *Le Peuple, La Sentinelle (LSE)*, whose node size decreased significantly. At the same time, the links between the Swiss newspapers and especially ATS-SDA become more pronounced, a development which even intensifies in the decade 1990-2000. It would be interesting to see if this co-occurred with a topic change in the articles from international to national news, to which the topic modelling from *impresso* could provide more insights. Throughout all time spans displayed in Figure 5.13, some Swiss newspapers (VHT, OIZ, SMZ, LCE) stayed almost unconnected to the network. Our guess is that they did not attribute agency content to the originators, to be determined with the help of the *impresso* text reuse clusters.

While the network density could already be discerned in the different network displays in Figure 5.13, its mathematical description gives a more exact estimation. For a (simple) bipartite graph $G_{\text{bipartite}}$, it is defined as follows:

$$\text{density}(G_{\text{bipartite}}) = \frac{|E|}{|N_0| \cdot |N_1|} = \frac{|\text{ existing edges }|}{|\text{ possible edges }|},$$

where $|.|$ denotes the absolute number of the variable it encloses, with E being the set of edges/links, N_0 the first node set and N_1 the second node set, and $N_0 \cap N_1 = \varnothing$ due to the bipartiteness.

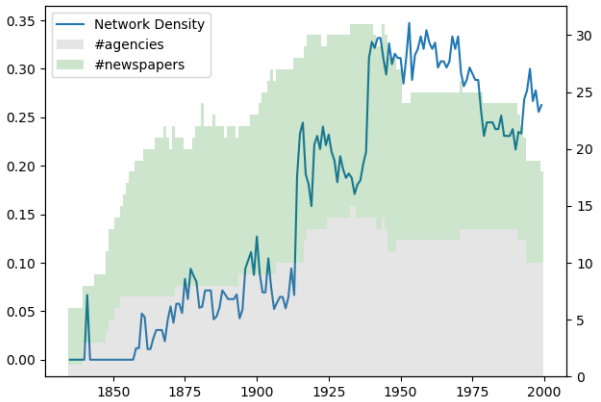


FIGURE 5.14

The yearly density of the network (left y-axis), plotted against the number of news agencies and newspapers in each year (right y-axis). A link between an agency and a newspaper was only considered if the newspaper cited the agency at least 24 times in the respective year (i.e. twice a month on average).

Figure 5.14 plots the network density per year[5], as well as the number of agencies and newspapers which existed in the respective year in the corpus. In general, there is no obvious relation between the network density and the number of newspapers and/or agencies in the network, so the development of the density can be interpreted independently.

The low number of references to agencies in the 19th century can also be observed here, although newspapers seem to start to cite agencies occasionally around 1860. Drastic increases in network density occur at the beginning of World War I in 1914 and again in 1939 for World War II, which could be caused by newspapers' desire to give a broad coverage of international news during wartime. Another hypothesis

[5]The multigraph of the agency network was reduced to a simple graph by combining all edges which occur between two nodes within a year to a (weighted) edge, so that we can use the definition of network density for a simple bipartite graph.

would be that newspapers more frequently disclosed the source of their information, or that they increased their share of international news, which often came from news agencies. Again, the methods of text reuse and topic modelling could offer further insights. The newspapers' concentration on a few agencies around the 1970s, which Figure 5.13 already hinted at, can be confirmed with this figure, as the network density decreases significantly during this time. However, the rise of density in the 1990s has no obvious explanation and would require a deeper inspection of the corpus.

5.3 Case Study: News Agencies in Luxembourg during the German Occupation in 1940-1944

We finish the analysis with a look into the situation of news agencies in Luxembourg during the Second World War.

Luxembourg was occupied by Nazi German troupes on the 10th of May 1940. Soon after the arrival of the Nazis, it became clear that the occupying power strove not only to ingest the country's lands, but also its people into the German nation – the goal was a total "Germanisation" of Luxembourg (Dostert 2003). For this, the French language was banned from public life on the 6th of August 1940, even forcing people to change their names to more German-sounding equivalents (ibid.).

For the originally very diverse Luxembourgish press landscape, this had severe consequences. All newspapers publishing in French were forced to cease their activities within a few weeks, and many German-speaking newspapers fell victim to the same fate eventually, such as the *Echternacher Anzeiger* on the 31st of December 1940 and the *Obermosel-Zeitung*, which existed until the end of 1941 (Bibliothèque nationale du Luxembourg 2021; Hilgert 2004). In October 1940, the Nazi administration took over the two daily newspapers *Luxemburger Wort* and *Escher Tageblatt*, which subsequently published content from collaborating Luxembourgish or German journalists (Hilgert 2004). By the end of 1941, those two newspapers and the party newspaper *Nationalblatt* were the only newspapers left on the Luxembourgish market (although the resistance also published several newspapers, albeit not as regularly (ibid.)).

Which impact did these extreme changes have on the use of news agencies in Luxembourg? Figures 5.15 and 5.16 provide insights, showing the development of agency citations in the French- and German-speaking newspapers respectively, the former for the years 1930-1949, and the latter on a monthly basis for the year 1940. According to those figures, references to agencies existed sparsely in the French-speaking newspapers before the war but saw a drastic increase in 1939, the share of articles with attributed agency content going up to 30% in 1940. Mainly, Havas and Reuters were used, while the Nazi-German DNB did not play a significant role. In May 1940, the citations stop abruptly (see Figure 5.16), mirroring the invasion of Germany and the subsequent publishing stop of French content.

The German-speaking newspapers exhibited a share of ca. 10% of articles with explicit agency content in the years before the war. We can see a varied mix of agencies, such as the German agencies Wolff, Europapress, and also DNB, the French Havas and the American UP-UPI. The *unk* class combines inter alia references to the French *Agence Fournier*, the German *Telunion* and the Belgian agency *Belga*. In the first few years of the DNB, the German-speaking newspapers gave the DNB much room, but from 1936 on, they referred to it less, while simultaneously increasing the content of Havas. This hints at a preference for French news and possibly cautious handling of information coming from the Nazis. A more detailed content analysis could find more robust answers to this idea.

At the moment of the German occupation, the use of agency content drastically changed, also for the German-speaking newspapers (see Figure 5.16). After the invasion in May 1940, the DNB remained as almost the only cited agency. However, also those citations stop abruptly after 1941. A look at Figure 5.17, which displays the distribution of the newspapers the agency mentions appeared in, gives answers: In

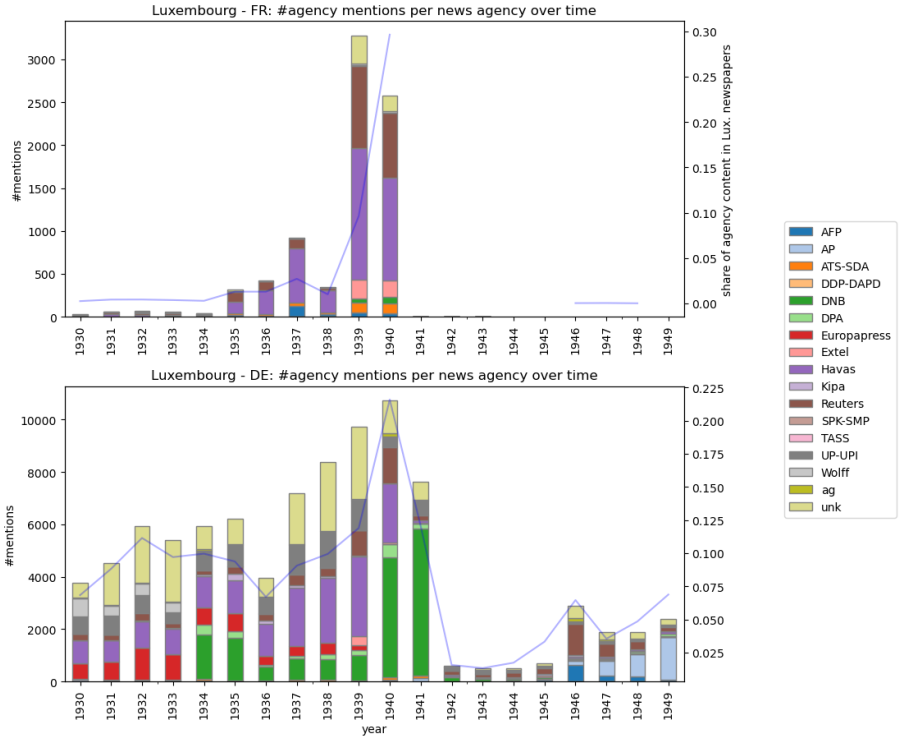


FIGURE 5.15
The development of agency mentions in Luxembourg in the years 1930-1949, split by language (French above, German below). The bars show the distribution of the different agency mentions (left y-axis), while the blue line indicates the general share of articles with agency mentions in the Luxembourgish corpus (right y-axis).

1941, almost all agency mentions came from the *Obermosel-Zeitung*, which had not been taken over by the Nazis until it was closed at the end of 1941. So apparently, with the enforced conformity of the two newspapers *Luxemburger Wort* and *Escher Tageblatt*, the practice of providing the source of news was given up. Figure 5.18 confirms this hypothesis, detailing the development of agency mentions in the *Luxemburger Wort* for the year 1940: Directly after the occupation, the newspaper still referenced the DNB (the mentions of DPA are false classifications, since the agency was founded in 1949), but the mentions disappear in October. This is incidentally also the month when the Nazi administration brought the newspaper into their control.

After the end of the war, the Luxembourgish press only recovered slowly from the suppression of the Nazis, some newspapers did not return at all (Hilgert 2004). Figure 5.15 indicates that German-speaking newspapers started to cite news agencies again in 1946, albeit not on the level as before the war.

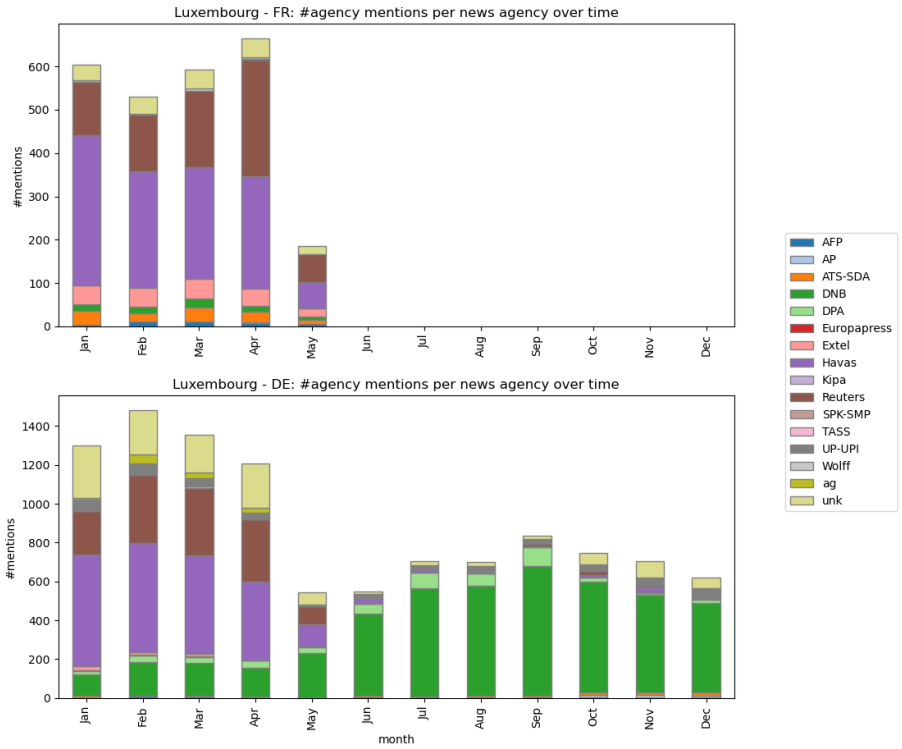


FIGURE 5.16
The development of agency mentions for the year 1940 in Luxembourg, split by language (French above, German below).

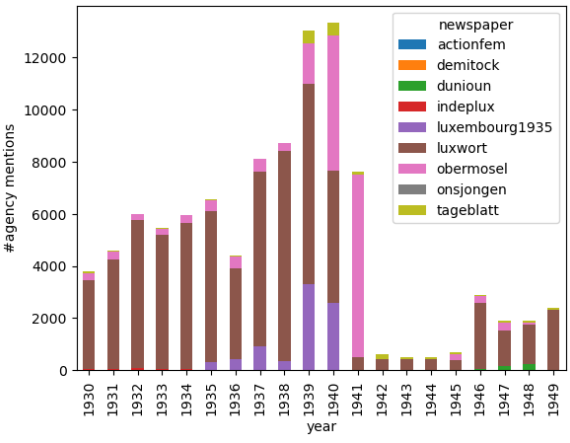


FIGURE 5.17
The development of agency mentions in the years 1930-1949 in Luxembourg, split by the newspaper they were cited in.

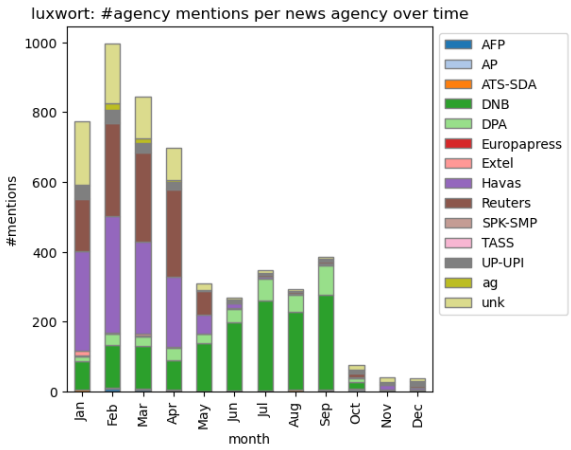


FIGURE 5.18
The development of agency mentions in the months of 1940 for the newspaper *Luxemburger Wort*.

6 Discussion and Outlook

To conclude the thesis, we give a short summary of the different steps we performed during the project, including a discussion of their respective strengths and limitations, and how the latter could be addressed in the future. We end by outlining research ideas based on the results of this thesis and their combination with earlier work on the *impresso* corpus, namely the text reuse clusters and topic modelling.

Recap and Discussion. Due to the fact that we did not dispose of news agency releases to trace agency content in the *impresso* corpus, we chose to focus on the newspaper articles where agencies were credited explicitly. This provides a good basis and can subsequently be enlarged with the help of text reuse clusters. Using the *impresso* app to query for agencies by their name, we assembled a raw corpus of 2,814,382 articles with potential agency content. From this, we sampled 1,610 articles for the annotation, adopting a sampling strategy which was uniform across the decades 1840-1990 and stratified over news agencies and newspapers. In retrospect, it would have been useful to additionally sample uniformly over languages, as the German language is underrepresented in the *impresso* corpus and thus also in the sampled dataset. The subsequent annotation campaign lasted around one month, where four annotators annotated the 1,610 articles, resulting in 1,976 news agency and *articleauthor* annotations. The moderate results of the inter-annotator agreement made a post-processing step necessary, which could rectify the most pressing discrepancies relevant to this project, such as missed annotations. The final dataset featured 22 different agency classes, with some smaller agencies subsumed in the generic *unk* class. Although we controlled this class regarding falsely classified agencies, we later found a French instance of SPK-SMP (the only one) under *unk*, an error which might have contributed to the absence of French SPK-SMP classifications at inference time. Aside from this error, we assume that the dataset includes all major news agencies in the *impresso* corpus, as we based the selection of news agencies both on literature and the inspection of several newspapers in *impresso* for all relevant decades. Considering the distribution of agency mentions across time, many of the sampled articles for the 19th century proved to be false positives (i.e. articles falsely assumed to have an agency mention). We suppose that this is due to an overall low number of agency citations in the *impresso* corpus in the 19th century. Therefore, the final dataset does not contain much data for this time, which might make it difficult for a classifier to detect agencies in this time period. The construction of an additional dataset which concentrates on the 19th century, maybe with the help of the agency mentions found by the classifier of this project, could address this issue. However, one might need to go through a lot of articles to find a sufficient amount of agency mentions, since the quality assessment of the classifications during inference showed that both the French and German models outputted many false positives in the 19th century.

Next, we strove to find a reliable machine learning classifier to detect agencies in the *impresso* corpus. The choice for the architecture fell on the transformer-based model BERT, since it performs well for named entity recognition and text classification tasks, is stable and exists in many versions pre-trained on various languages, including in-domain data (i.e. historical newspapers). During our experiments, we tested different model configurations, which mainly varied in their pre-training data, using generic vs. in-domain data, and German, French or multilingual training sets. We kept all hyperparameters fixed except the

maximum sequence length, where we investigated the changes in performance for four different sequence lengths. Compared to the general results for historical NER, the performance of the best models for the agency detection task was satisfying, with F-scores around 0.8. Although the F-score of 0.73 from the lookup baseline relativised the results a bit, the BERT-based models still achieved higher precision scores, thus providing a "cleaner" classification with fewer false positives. During training, we observed more unstable results for German as well as for both languages in the 19th century, errors which were probably caused by the imbalances in the fine-tuning dataset. Surprisingly, the models pre-trained on historical data did not always perform better than their counterparts pre-trained on generic data, which might be due to different pre-training schemes and the missing hyperparameter tuning on our part. Although the multilingual models profited from the simultaneous fine-tuning on both the French and German datasets, they still underperformed compared to the models solely trained on one of the two languages. Thus, we chose two different models for the inference on the *impresso* corpus, one for French (French Europeana BERT) and one for German (German BERT). While we expected the models to perform robustly for unseen (not noisy) data, we did not test all possible routes to improve their performance. For example, future experiments could incorporate the *impresso* (in-domain) word embeddings, make an exhaustive hyperparameter search and augment the training data with the help of the *impresso* text reuse clusters. Additionally, an article-level classification as opposed to the performed sentence-level classification could train the classifier to learn even more linguistic characteristics and structural aspects of agency-based articles, which might enable the classifier to find agency content without explicit mention of an agency. However, it would need to be investigated how well such a classifier can perform, as it is unclear whether agency articles are clearly distinguishable from the rest solely on the basis of linguistic features, and partly due to unclear article boundaries because of imprecise OLR results in *impresso*.

Thanks to the contribution of Emanuela Boros and Maud Ehrmann, we were able to execute the inference process on the 25 million articles in the *impresso* corpus during the (limited) time of this thesis. By making use of two different parallelization libraries in Python, the time frame for the application of the two chosen agency detection models on the corpus could be reduced to two weeks. The subsequent quality assessment of the detected agency mentions showed that the classifications in the 20th century were generally reliable, with a slight decrease in performance for the latter decades in the German part of the corpus. All results until around 1890 should be treated with caution, though, as they contain many false positives. Here, the consequences of error propagation become visible: During the construction of the training dataset, the annotation revealed a low number of true agency mentions in the 19th century, producing a skew towards 20th-century data. Thus, the training of the agency detection model was imbalanced, causing it to perform less steadily on articles from the 19th century. Altogether, we have to be aware that errors accumulate across the different processing steps and equally arise due to the variety and noisiness caused by the big number of documents and the long time span of the data. However, the quality assessment also suggested that most of the classifications are correct and provide a solid basis for comprehensive analyses.

We conducted a first analysis on the share of agency content in the *impresso* corpus, finding that agencies were mainly cited in the 20th century, with the share of articles with agency mentions increasing significantly at the beginning of World War I and II respectively. Taking the precision of the agency detection models into account, we estimated the proportion of articles with agency citations to lie roughly between 13 and 22% after 1940. In the 1940s, references to news agencies seemed to become a widespread custom, although some newspapers already started to cite agencies systematically in the 1910s, and one (*luxwort*) even in the 1890s. The percentages of explicitly credited agency content are comparable to the results reported in the literature for the 21st century, although many contemporary studies dispose of the original agency releases, allowing them to make more precise estimates. After the *impresso* text reuse clusters are integrated with the analysis, we expect the share of articles with agency content to rise significantly, especially since some newspapers such as the *Confedéré* very rarely credited agencies (see Figure 5.13). Boumans (2018) found that apart from newspapers which did not cite agencies at all, newspapers credited

agencies in 70-94% of the time they used agency content. So we expect to detect additional content through text reuse for newspapers which usually refer to agencies as well.

Compared to the research on the influence of news agencies on newspapers so far, this study is the first one, to our knowledge, which has the resources to trace agencies for nearly two centuries – since their foundation – in a considerable number of newspapers. This opens up many research possibilities, of which the scope of this thesis, and to a certain extent also our limited historical knowledge, only allowed us to address a few. Still, we were able to get a broad picture of news agency content in the *impresso* corpus, always putting our findings into context, discussing possible explanations for unexpected results and outlining research directions to find answers to open questions and hypotheses. We found that the effects of the news agency cartel between 1859 and 1933 were also visible for the news market in Switzerland, with Havas in a dominant position, followed by Wolff and the Swiss ATS-SDA, founded in 1894 as a counterweight to the "big" internationally operating agencies. Reuters only slowly appeared on the Swiss market, in contrast to Luxembourg, where it was already cited by the end of the 19th century. After the end of the cartel, the agency landscape became more diverse in both countries, especially during and shortly after World War II, until it concentrated on a few agencies again around the 1970s. Next to differences over time, we also observed variation with regard to languages, as the German-speaking newspapers in both Switzerland and Luxembourg relied on a more diverse set of agencies compared to their French-speaking competition. In a case study on news agencies in Luxembourg during the German occupation (1940-44), we discovered that in parallel to the enforced conformity of newspapers by the Nazis, the lively use of news agencies in Luxembourg was equally reduced, first concentrating on the Nazi agency DNB, and with the takeover of newspapers by the Nazi administration, the practice of referencing news agencies ceased completely. As far as we know, this case study is the first to look at news agencies during the German occupation of Luxembourg, thus contributing to the historical research on this period.

Future Work. Having finished the inference process on the *impresso* corpus enabled us to conduct a first analysis on the detected agency mentions. In the second step, this data needs to be ingested in the Solr system, which holds the actual text of the articles, as well as all data enrichments such as text reuse clusters or topic modelling developed during the first *impresso* project. This process was not finished in time in order to be exploited for this thesis, but pairing the detected agency mentions with all the additional information will open up yet another bundle of research possibilities.

Which news did agencies usually provide to newspapers? Did they have certain topics they specialised in? Did this change over the years? – Such questions can be answered with the help of topic modelling. Our analysis showed that while Swiss newspapers cited several different news agencies in the first half of the 20th century, a concentration on a few agencies such as ATS-SDA occurred in the second half, and newspapers used a lot of agency content from a few agencies. We would like to check if this was accompanied by an enlargement of topics as well, or if they relied mostly on agency content for international news, like the study by UNESCO (1953) suggests. Moreover, we have the hypothesis that in times of war, newspapers increased the amount of international news and relied heavily on agencies for war reports. A study of the topics covered during this time could shed light on this idea as well.

The text reuse clusters give way to an even broader field of possible research directions. First, they complement the research we already did on the amount of agency content in newspapers, thus giving a more complete notion of the leverage of news agencies to shape the news world in the past, and opening the possibility to compare it to studies on the situation in the 21st century. Additionally, some of the questions which motivated this project could finally be answered, such as the extent to which journalists relied on agency content to produce their stories, or if agency releases were printed as simple verbatim (copy and paste) or with rephrasing in the newspapers. Comparisons between different newspapers could show which newspapers systematically credited agencies and which did not, or which newspapers used the same agency releases, possibly providing information on newspapers' political orientation. Furthermore,

as agencies had contracts with each other to exchange their news, it could be interesting to see whether cases exist where the same text was printed under different agency names in different newspapers. We also have numbers on the amount of news reports the major agencies published, so another line of questions could go in the direction of the extent to which those reports were actually printed in the newspapers and investigate if this stands in relation to the success of a news agency. Last but not least, studies on the flow of news (e.g. as conducted by Salmi et al. 2020) become possible, for example, analysing which agency releases were the most influential, how long news usually took to spread or which agency releases became viral. The position of newspapers in the news network could also be studied, such as which newspapers generally printed news the fastest and which newspapers copied from other newspapers, maybe even without subscribing to agency services themselves.

Furthermore, the analysis conducted in the last chapter leaves open questions and research ideas which can be expanded on. Many of them require the topic and text reuse data and thus were already discussed above. Additionally, the method of network analysis opens research opportunities by providing another view of the workings of the news ecosystem as a whole. By projecting the bipartite network from the previous chapter to a newspaper-newspaper network on the one hand, or an agency-agency network on the other, we can address questions such as which agencies deliver to the same newspapers or which never serve the same, if there are clusters within the network, and if the location of agencies and newspapers, as well as language and country boundaries, play a significant role. Another potential of the corpus where we only scratched the surface is its multi-nationality, which will be even more pronounced when the corpus is expanded to include more media and additional countries during the second *impresso* project. Until now, we only roughly compared the situations in Switzerland and Luxembourg, but more and deeper comparative studies are possible. For example, in our case study, we examined the situation of news agencies in Luxembourg during the Second World War. It would be informative to compare the pre-war years as well as the years during the war to the situation in Switzerland, a country which – in contrast to Luxembourg – managed to stay neutral. How did these political differences translate to the news world? And can we observe differences during this time even within Switzerland, across language borders?

Apart from this, the ingestion of the detected agency mentions in the *impresso* app will allow for all *impresso* users to filter on agency content. This could for example be used for qualitative historical research, or complement the study of the history of a single newspaper or agency.

Finally, the quality of the agency classification could be revisited as well. Objectives could be a performance improvement or the inclusion of more languages (e.g. English, as it is introduced in the second *impresso* project). A classifier trained to detect agency content on an article level could check whether agency reports were written in a distinct style, as suggested by Boyd-Barrett and M. Palmer (1981, p. 39f). This would open the door to analyses with a linguistic focus.

Conclusion. This thesis contributed to the research on news agencies in three main points: (1) The construction of a dataset with 1,976 annotated agency mentions, (2) the training of a series of BERT-based agency detection models, of which the two best models (one for French and one for German) were chosen to be applied to the *impresso* corpus which contains ca. 25 million articles, and (3) the analysis of the 2.4 million articles with detected agency mentions. The analysis provided an overview of the role of news agencies in the newspaper landscape of Switzerland and Luxembourg and introduced ways to explore the data further. All in all, although the different aspects of the thesis are grounded in the literature, this project is unique in its approach to detecting agencies in a multilingual and multinational corpus across almost two centuries. In contrast to most studies on the role of news agencies in contemporary media, we could not rely on the original agency releases for our project, a shortcoming which can partly be compensated using the *impresso* text reuse clusters. The incorporation of the latter was not (yet) possible due to the time constraints of the thesis, but the findings we were able to get even without the text reuse are already very informative, promising exciting possibilities for future work.

Bibliography

Akbik, Alan, Duncan Blythe and Roland Vollgraf (Aug. 2018). 'Contextual String Embeddings for Sequence Labeling'. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 1638–1649.

Ba, Jimmy Lei, Jamie Ryan Kiros and Geoffrey E. Hinton (July 2016). *Layer Normalization*. arXiv:1607.06450 [cs, stat]. DOI: 10.48550/arXiv.1607.06450.

Bibliothèque nationale du Luxembourg (2021). *Obermosel-Zeitung*. https://persist.lu/ark:70795/nhbpwsjbxd.

Bojanowski, Piotr et al. (June 2017). *Enriching Word Vectors with Subword Information*. arXiv:1607.04606 [cs]. DOI: 10.48550/arXiv.1607.04606.

Boros, Emanuela et al. (Sept. 2020). 'Robust Named Entity Recognition and Linking on Historical Multilingual Documents'. In: vol. 2696. Issue: Paper 171. CEUR-WS Working Notes. DOI: 10.5281/zenodo.4068074.

Boumans, Jelle (2018). 'The Agency Makes the (Online) News World Go Round: The Impact of News Agency Content on Print and Online News'. In: *International Journal of Communication* 12, p. 22.

Boyd-Barrett, Oliver (Feb. 2000). 'National and International News Agencies: Issues of Crisis and Realignment'. In: *Gazette (Leiden, Netherlands)* 62.1. Publisher: SAGE Publications, pp. 5–18. ISSN: 0016-5492. DOI: 10.1177/0016549200062001001.

Boyd-Barrett, Oliver and Michael Palmer (1981). *Le trafic des nouvelles. Les agences mondiales de l'information*. Paris: Alain Moreau.

British Library Labs (2016). *Digitised Books. c. 1510 - c. 1946. JSON (OCR derived text)*. DOI: 10.21250/DB14.

Chan, Branden et al. (2019). *German BERT | State of the Art Language Model for German NLP*. https://www.deepset.ai/german-bert.

Chen, Shijie, Yu Zhang and Qiang Yang (Sept. 2021). *Multi-Task Learning in Natural Language Processing: An Overview*. arXiv:2109.09138 [cs]. DOI: 10.48550/arXiv.2109.09138.

Cho, Kyunghyun et al. (Oct. 2014). *On the Properties of Neural Machine Translation: Encoder-Decoder Approaches*. arXiv:1409.1259 [cs, stat]. DOI: 10.48550/arXiv.1409.1259.

Clavien, Alain (2010). *Grandeurs et misères de la presse politique: le match, Gazette de Lausanne, Journal de Genève*. GRHIC. Lausanne, Suisse: Antipodes. ISBN: 978-2-940146-99-4.

Collobert, Ronan et al. (Mar. 2011). *Natural Language Processing (almost) from Scratch*. arXiv:1103.0398 [cs]. DOI: 10.48550/arXiv.1103.0398.

Conneau, Alexis, Kartikay Khandelwal et al. (Nov. 2019). *Unsupervised Cross-lingual Representation Learning at Scale*. DOI: 10.48550/arXiv.1911.02116.

Conneau, Alexis and Guillaume Lample (2019). 'Cross-lingual Language Model Pretraining'. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc.

Cortes, Corinna and Vladimir Vapnik (Sept. 1995). 'Support-Vector Networks'. In: *Machine Language* 20.3, pp. 273–297. ISSN: 0885-6125. DOI: 10.1023/A:1022627411411.

Czarniawska-Joerges, Barbara (Jan. 2011). *Cyberfactories: How News Agencies Produce News*. Google-Books-ID: QDGW43rRFJ8C. Edward Elgar Publishing. ISBN: 978-0-85793-913-5.

Deshpande, Ameet, Partha Talukdar and Karthik Narasimhan (May 2022). *When is BERT Multilingual? Isolating Crucial Ingredients for Cross-lingual Transfer*. arXiv:2110.14782 [cs]. DOI: 10.48550/arXiv.2110.14782.

Devlin, Jacob et al. (May 2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv:1810.04805 [cs]. DOI: 10.48550/arXiv.1810.04805.

Dos Santos, Cícero Nogueira and Bianca Zadrozny (June 2014). 'Learning character-level representations for part-of-speech tagging'. In: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. ICML'14. Beijing, China: Journal of Machine Learning Research, pp. II–1818–II–1826.

Dostert, Paul (Dec. 2003). 'Luxemburg unter deutscher Besatzung 1940-45'. de. In: *Gedenkstättenrundbrief* 116, pp. 33–43.

Ehrmann, Maud, Ahmed Hamdi et al. (June 2023). 'Named Entity Recognition and Classification in Historical Documents: A Survey'. In: *ACM Computing Surveys*. Just Accepted. ISSN: 0360-0300. DOI: 10.1145/3604931.

Ehrmann, Maud, Matteo Romanello, Simon Clematide et al. (May 2020). 'Language resources for historical newspapers: the *Impresso* collection'. In: *Proceedings of the 12th language resources and evaluation conference*. Marseille, France: European Language Resources Association, pp. 958–968.

Ehrmann, Maud, Matteo Romanello, Alex Flückiger et al. (Sept. 2020). 'Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers'. In: *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*. Ed. by Linda Cappellato et al. Vol. 2696. CEUR Workshop Proceedings. ISSN: 1613-0073. Thessaloniki, Greece: CEUR.

Ehrmann, Maud, Matteo Romanello, Sven Najem-Meyer et al. (Aug. 2022). 'Extended Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents'. In: Bologna, Italy. DOI: 10.5281/zenodo.6979577.

'Telegraphenbureau.' (1894). In: *Brockhaus Konversationslexikon*. Ed. by F. A. Brockhaus. 14th ed. Vol. 15. Leipzig, Berlin, Wien, p. 668.

Forde, Susan and Jane Johnston (Feb. 2013). 'The News Triumvirate'. In: *Journalism Studies* 14.1, pp. 113–129. ISSN: 1461-670X. DOI: 10.1080/1461670X.2012.679859.

Groth, Otto (1928). *Die Zeitung: Ein System der Zeitungskunde*. Mannheim, Germany: J. Bensheimer.

He, Jianming (1996). *Die Nachrichtenagenturen in Deutschland: Geschichte und Gegenwart*. Vol. 58. European university studies. Series XL, Communications. Frankfurt am Main ; New York: P. Lang. ISBN: 978-3-631-49394-6.

He, Kaiming et al. (June 2016). 'Deep Residual Learning for Image Recognition'. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 1063-6919, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

Hendrycks, Dan and Kevin Gimpel (June 2016). *Gaussian Error Linear Units (GELUs)*. arXiv:1606.08415 [cs] version: 1. DOI: 10.48550/arXiv.1606.08415.

Hilgert, Romain (Oct. 2004). *Zeitungen in Luxemburg 1704-2004*. Service information et presse du gouvernement luxembourgeois.

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). 'Long Short-Term Memory'. In: *Neural Computation* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735.

Institut international de la presse (1953). *La circulation des informations: étude menée avec la collaboration de directeurs de journaux et d'agences de presse ainsi que les correspondants à l'étranger dans dix pays*. Zürich: Institut international de presse.

Klie, Jan-Christoph et al. (Aug. 2018). 'The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation'. In: *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. Santa Fe, New Mexico: Association for Computational Linguistics, pp. 5–9.

Krippendorff, Klaus (2004). *Content Analysis: An Introduction to Its Methodology*. 2nd ed. Sage Publications.

Lafferty, J., A. McCallum and Fernando Pereira (June 2001). 'Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data'. In: *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–289.

Lample, Guillaume et al. (Apr. 2016). *Neural Architectures for Named Entity Recognition*. arXiv:1603.01360 [cs]. DOI: 10.48550/arXiv.1603.01360.

Lan, Zhenzhong et al. (Feb. 2020). *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*. arXiv:1909.11942 [cs]. DOI: 10.48550/arXiv.1909.11942.

Lauriola, Ivano, Alberto Lavelli and Fabio Aiolli (Jan. 2022). 'An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools'. In: *Neurocomputing* 470, pp. 443–456. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2021.05.103.

Levenshtein, V. (1965). 'Binary codes capable of correcting deletions, insertions, and reversals'. In: *Soviet physics. Doklady*.

Lewis, Justin, Andrew Williams and Bob Franklin (Feb. 2008). 'A Compromised Fourth Estate?' In: *Journalism Studies* 9.1. Publisher: Routledge, pp. 1–20. ISSN: 1461-670X. DOI: 10.1080/14616700701767974.

Liu, Yinhan et al. (July 2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv:1907.11692 [cs]. DOI: 10.48550/arXiv.1907.11692.

Loshchilov, Ilya and Frank Hutter (Jan. 2019). *Decoupled Weight Decay Regularization*. arXiv:1711.05101 [cs, math]. DOI: 10.48550/arXiv.1711.05101.

Martin, Louis et al. (2020). 'CamemBERT: a Tasty French Language Model'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. arXiv:1911.03894 [cs], pp. 7203–7219. DOI: 10.18653/v1/2020.acl-main.645.

Meier, Peter (Sept. 2010). *Nachrichtenagenturen*. https://hls-dhs-dss.ch/articles/010466/2010-09-02/.

Meyer, ed. (1909). *Telegraphenbureaus.* http://www.zeno.org/nid/20007569742. Leipzig.

Mikolov, Tomas, Kai Chen et al. (Sept. 2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv:1301.3781 [cs]. DOI: 10.48550/arXiv.1301.3781.

Mikolov, Tomas, Wen-tau Yih and Geoffrey Zweig (June 2013). 'Linguistic Regularities in Continuous Space Word Representations'. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, pp. 746–751.

Minaee, Shervin et al. (Apr. 2021). 'Deep Learning–based Text Classification: A Comprehensive Review'. In: *ACM Computing Surveys* 54.3, 62:1–62:40. ISSN: 0360-0300. DOI: 10.1145/3439726.

Neudecker, Clemens and Apostolos Antonacopoulos (Apr. 2016). 'Making Europe's Historical Newspapers Searchable'. In: *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pp. 405–410. DOI: 10.1109/DAS.2016.83.

Nicholls, Tom (Sept. 2019). 'Detecting Textual Reuse in News Stories, At Scale'. In: *International Journal of Communication* 13. Number: 0, p. 25. ISSN: 1932-8036.

Nothman, Joel, James R. Curran and Tara Murphy (Dec. 2008). 'Transforming Wikipedia into Named Entity Training Data'. In: *Proceedings of the Australasian Language Technology Association Workshop 2008*. Hobart, Australia, pp. 124–132.

Palmer, Michael B. (2019). *International News Agencies: A History*. Cham: Springer International Publishing. ISBN: 978-3-030-31177-3 978-3-030-31178-0. DOI: 10.1007/978-3-030-31178-0.

Paszke, Adam et al. (Dec. 2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. arXiv:1912.01703 [cs, stat]. DOI: 10.48550/arXiv.1912.01703.

Pennington, Jeffrey, Richard Socher and Christopher Manning (Oct. 2014). 'GloVe: Global Vectors for Word Representation'. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural*

Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

Peters, Matthew E. et al. (Mar. 2018). *Deep contextualized word representations*. arXiv:1802.05365 [cs]. DOI: 10.48550/arXiv.1802.05365.

Rabiner, L. and B. Juang (Jan. 1986). 'An introduction to hidden Markov models'. In: *IEEE ASSP Magazine* 3.1, pp. 4–16. ISSN: 1558-1284. DOI: 10.1109/MASSP.1986.1165342.

Radford, Alec and Karthik Narasimhan (2018). *Improving Language Understanding by Generative Pre-Training*. https://api.semanticscholar.org/CorpusID:49313245.

Rantanen, T. (1990). *Foreign news in imperial Russia: The relationship between international and Russian news agencies, 1856-1914*. https://api.semanticscholar.org/CorpusID:150637206.

Rantanen, Terhi (Mar. 2019). *News Agencies from Telegraph Bureaus to Cyberfactories*. ISBN: 9780190228613. DOI: 10.1093/acrefore/9780190228613.013.843.

Ruder, Sebastian et al. (June 2019). 'Transfer Learning in Natural Language Processing'. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 15–18. DOI: 10.18653/v1/N19-5004.

Sadvilkar, Nipun and Mark Neumann (Oct. 2020). *PySBD: Pragmatic Sentence Boundary Disambiguation*. arXiv:2010.09657 [cs]. DOI: 10.48550/arXiv.2010.09657.

Salmi, Hannu et al. (Sept. 2020). 'The Reuse of Texts in Finnish Newspapers and Journals, 1771– 1920: A Digital Humanities Perspective'. In: *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 54.1. Publisher: Routledge, pp. 14–28. ISSN: 0161-5440. DOI: 10.1080/01615440.2020.1803166.

Sanh, Victor et al. (Feb. 2020). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv:1910.01108 [cs]. DOI: 10.48550/arXiv.1910.01108.

Schramm, Wilbur (1959). *One day in the world's press: fourteen great newspapers on a day of crisis Nov. 2, 1956. With translations and facsimiles reproductions*. Stanford: Stanford Univ. Press.

Schwarzlose, Richard Allen (1989). *The Nation's Newsbrokers: The formative years, from pretelegraph to 1865*. Northwestern University Press. ISBN: 978-0-8101-0818-9.

Schweter, Stefan (Nov. 2020). *Europeana BERT and ELECTRA models*. DOI: 10.5281/ZENODO.4275044.

Schweter, Stefan and Alan Akbik (May 2021). *FLERT: Document-Level Features for Named Entity Recognition*. arXiv:2011.06993 [cs]. DOI: 10.48550/arXiv.2011.06993.

Schweter, Stefan, Luisa März et al. (May 2022). *hmBERT: Historical Multilingual Language Models for Named Entity Recognition*. DOI: 10.48550/arXiv.2205.15575.

Shrivastava, K. M. (2007). *News Agencies from Pigeon to Internet*. Sterling Publishers Pvt. Ltd. ISBN: 978-1-932705-67-6.

Silberstein-Loeb, Jonathan (2014). *The International Distribution of News: The Associated Press, Press Association, and Reuters, 1848–1947*. Cambridge Studies in the Emergence of Global Enterprise. Cambridge: Cambridge University Press. ISBN: 978-1-107-03364-1. DOI: 10.1017/CBO9781139522489.

Suarez, Pedro Ortiz (July 2019). *Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures*. DOI: 10.14618/IDS-PUB-9021.

Sun, Yu et al. (Nov. 2019). *ERNIE 2.0: A Continual Pre-training Framework for Language Understanding*. arXiv:1907.12412 [cs]. DOI: 10.48550/arXiv.1907.12412.

UNESCO (1953). *News agencies: their structure and operation*. Paris: UNESCO.

Vaswani, Ashish et al. (2017). *Attention Is All You Need*. Publisher: arXiv Version Number: 6. DOI: 10.48550/arXiv.1706.03762.

Vogler, Daniel, Linards Udris and Mark Eisenegger (Aug. 2020). 'Measuring Media Content Concentration at a Large Scale Using Automated Text Comparisons'. In: *Journalism Studies* 21.11. Publisher: Routledge, pp. 1459–1478. ISSN: 1461-670X. DOI: 10.1080/1461670X.2020.1761865.

Wang, Xinyu et al. (Aug. 2021). 'Automated Concatenation of Embeddings for Structured Prediction'. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 2643–2660. DOI: 10.18653/v1/2021.acl-long.206.

Welbers, Kasper et al. (Feb. 2018). 'A Gatekeeper among Gatekeepers: News agency influence in print and online newspapers in the Netherlands'. In: *Journalism Studies* 19.3, pp. 315–333. ISSN: 1461-670X, 1469-9699. DOI: 10.1080/1461670X.2016.1190663.

Whang, Steven Euijong et al. (July 2023). 'Data collection and quality challenges in deep learning: a data-centric AI perspective'. In: *The VLDB Journal* 32.4, pp. 791–813. ISSN: 0949-877X. DOI: 10.1007/s00778-022-00775-9.

Wikimedia Foundation (2023). *Wikipedia, the free encyclopedia*.

Windlinger, Andreas (Oct. 2011). *Schweizerische Politische Korrespondenz (SPK)*. Historisches Lexikon der Schweiz (HLS). https://hls-dhs-dss.ch/articles/043156/2011-10-28/.

Wu, Shijie and Mark Dredze (Nov. 2019). 'Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT'. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 833–844. DOI: 10.18653/v1/D19-1077.

Yadav, Vikas and Steven Bethard (Oct. 2019). *A Survey on Recent Advances in Named Entity Recognition from Deep Learning models*. arXiv:1910.11470 [cs]. DOI: 10.48550/arXiv.1910.11470.

Yin, Wenpeng et al. (Feb. 2017). *Comparative Study of CNN and RNN for Natural Language Processing*. arXiv:1702.01923 [cs]. DOI: 10.48550/arXiv.1702.01923.

Zhou, Ce et al. (May 2023). *A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT*. arXiv:2302.09419 [cs]. DOI: 10.48550/arXiv.2302.09419.

Zhu, Yukun et al. (June 2015). *Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books*. arXiv:1506.06724 [cs]. DOI: 10.48550/arXiv.1506.06724.

A List of Newspapers in *impresso*

TABLE A.1
Newspapers and their Abbreviations from *impresso*, used in this project.

Begin of Table A.1 (Newspapers in *impresso*)

ID	**Newspaper Title**	***impresso* Start Year**	***impresso* End Year**
actionfem	L'Action féminine	1927	1940
armeteufel	Arme Teufel	1903	1929
avenirgdl	L'avenir	1868	1871
BDC	Bulletin des séances de la Constituante	1839	1839
BLB	Bündner Landbote	1846	1847
BNN	Bündner Nachrichten	1885	1892
buergerbeamten	Bürger- und Beamten-Zeitung	1898	1916
CDV	Courrier du Valais	1843	1857
CON	La Contrée	1902	1903
courriergdl	Courrier du Grand-Duché de Luxembourg	1844	1868
deletz1893	De Letzeburger	1893	1909
demitock	De Mitock e Wocheblad fir Jux an Zodi	1937	1940
DFS	"Wochenblatt für die vier löblichen Kantone Ury, Schwytz, Unterwalden und Zug"	1814	1849
diekwochen	Diekircher Wochenblatt	1841	1848
DLE	Der Landbote des freiburgischen Seebezirks	1909	1914
DTT	Die Tat	1935	1978
dunioun	D'Unio'n	1944	1948
DVF	Der Volksfreund	1879	1885
EDA	L'Écho des Alpes	1839	1844
EXP	L'Express	1738	2018
EZR	Der Erzähler	1806	1865
FCT	La FCTA	1955	1963
FZG	Freiburger Nachrichten	1865	2018
GAV	Gazette du Valais / Nouvelle gazette du Valais	1855	1922
GAZ	Gazette du Simplon	1842	1847
gazgrdlux	Gazette du Grand-Duché de Luxembourg	1878	1878
GDL	Gazette de Lausanne	1804	1991
HRV	Der helvetische Volksfreund	1799	1801
IMP	L'Impartial	1881	2018
indeplux	L'indépendance luxembourgeoise	1871	1934
JDF	Journal du canton de Fribourg	1830	1833

Continuation of Table A.1 (Newspapers in *impresso*)

ID	**Newspaper Title**	***impresso* Start Year**	***impresso* End Year**
JDG	Journal de Genève	1826	1998
JDV	Le Journal du Valais	1848	1848
kommmit	Komm mit mir!	1884	1884
LAB	Der Liberale Alpenbote	1848	1860
landwortbild	Luxemburger Land in Wort und Bild	1895	1895
LBP	Le Bien public	1879	1888
LCE	Confédéré	1861	2018
LCG	Le Confédéré de Fribourg	1848	1907
LCR	Le Chroniqueur	1854	1881
LCS	Le Courrier fribourgeois	1830	1830
LES	L'Essor	1906	2018
LLE	La Liberté	1871	2018
LLS	La lutte syndicale	1906	1998
LNF	Le Narrateur fribourgeois	1840	1855
LSE	"Le Peuple, La Sentinelle"	1890	1971
LSR	L'Observateur	1846	1848
LTF	La Tribune de Fribourg	1905	1905
lunion	L'Union	1860	1871
luxembourg1935	Luxembourg (1935)	1935	1940
luxland	d'Letzeburger Land	1954	2007
luxwort	Luxemburger Wort	1848	1950
luxzeit1844	Luxemburger Zeitung	1844	1845
luxzeit1858	Luxemburger Zeitung - Journal de Luxembourg	1858	1859
LVE	Le Véridique	1831	1833
MGS	Der Morgenstern	1842	1843
NTS	Neues Tagblatt aus der östlichen Schweiz	1856	1874
NZG	Neue Zuger Zeitung (II)	1846	1891
NZZ	Neue Zürcher Zeitung	1780	1950
obermosel	Obermosel-Zeitung	1881	1948
OIZ	Die Gewerkschaft	1901	1992
onsjongen	Ons Jongen	1944	1951
schmiede	Schmiede	1916	1919
SDT	Solidarité	1909	2001
SGZ	St.Galler Zeitung	1831	1881
SMZ	SMUV-Zeitung	1902	2004
SRT	Schweizerische Tag-Blätter	1798	1798
tageblatt	Escher Tageblatt	1913	1950
VHT	VHTL-Zeitung	1904	2004
volkfreu1869	Volksfreund	1890	1939
waechtersauer	Der Wächter an der Sauer	1849	1869
waeschfra	D'Wäschfra	1868	1884
WHD	Der Wahrheitsfreund	1835	1863
ZBT	Der Zugerbieter	1865	1868

B List of Queried News Agencies in *impresso*

TABLE B.1
Table containing details of queried news agencies.

Begin of Table B.1 (Queried news agencies)

Abbr.	News Agency	Country	Creation	End
AA	Athens News Agency	Greece	1905	
ACP	Agence centrale de presse	France	1951	1993
ADN	Allgemeiner Deutscher Nachrichtendienst	Germany (GDR)	1946	
Adnkronos	Adnkronos	Italy	1963	
AFP	Agence France-Presse	France	1944	
AGERPRESS	Agentia Romana de Presa	Romania	1889	
ANP	Algemeen Nederlands Persbureau	Netherlands	1934	
ANSA	Agenzia Nationale Stampa Associata	Italy	1945	
AP	Associated Press	USA	1848	
APA	Austria Press Agentur	Austria	1946	
ATS / SDA	Agence Telegraphique Suisse/ Schweizerische Depeschenagentur	Switzerland	1894	
BELGA	Agence télégraphique belge de Presse	Belgium	1920	
BelTA	Belarusian Telegraph Agency	Belarus	1918	
BTA	Bulgarska Telegrafitscheka Agentzia/Agence Bulgare	Bulgaria	1898	
CAPA	Chabalier & Associates Press Agency	France	1989	
CP	Canadian Press	Canada	1917	
CTK	Czechoslavenska Tiskova Kancelar/ Agence Ceteka	Czechoslovakia/Czech	1918	
DDP/DAPD	Deutscher Depeschendienst/Deutscher Auslands-Depeschendienst	Germany	1971	2013
DNB	Deutsches Nachrichtenbüro GmbH	Germany	1933	1945
DOMEI	Domei Tsushin/Domein News Agency	Japan	1936	1945
DPA	Deutsche Presse Agentur	Germany (GFR)	1949	
EFE	Agencia EFE, S.A.	Spain	1938	
EPD	Evangelischer Pressedienst	Germany	1918	
Europa Press	Europa Press	Spain	1953	

Continuation of Table B.1 (Queried news agencies)

Abbr.	News Agency	Country	Creation	End
Europapress	Europapress	Germany		
Extel	Exchange Telegraph Co. Ltd.	United Kingdom	1872	1993
Fides		Vatican	1927	
Fournier	Agence de Presse Fournier	France	1874	
HAVAS	Havas	France	1835	1940
INS	International News Service	USA	1909	1958
Interfax	Interfax News Agency	Russia	1989	
JIJI	Jiji Press	Japan	1945	
Kipa/Apic	Katholische internationale Presseagentur / Agence de presse internationale catholique	Switzerland	1917	2015
KKTK/ Korbureau	Telegraphen-Korrespondenz Bureau/ Kaiserlich und Königlich Telegraphen-Korrespondenz Bureau	Austria	1849	1945
KYODO	Kyodo Tsu Dhinsha/Kyodo News	Japan	1945	
LUSA	Lusa News Agency	Portugal	1986	
Meurisse	Agence Meurisse	France	1909	1937
MTI	Magyar Tavirati Iroda (bureau télégraphique hongrois)	Hungary	1881	
NCNA/ Xinhua	Xinhua (New China News Agency)	China (People's Republic of)	1937	
NTB	Norsk Telegrambyra P/S	Norway	1867	
Office-Correspondance	Agence Bullier/Office-Correspondance pour les journaux français et étrangers et pour les affaires en fonds publics à la Bourse de Paris	France	1830	1870
Ottomane (?)	Agence télégraphique ottomane	Europe	1909	1923
PA	The Press Association/PA Media	United Kingdom	1868	
PAP	Polska Agencja Prasowa	Poland	1944	
PTI	Press Trust of India	India	1947	
RB / RITZAUS	Ritzaus Bureau	Denmark	1866	
REUTERS	Reuters	United Kingdom	1851	
Rol	Agence Rol	France	1904	1937
ROSTA	Russische Telegraphenagentur	Russia	1918	1935
SAMACHAR	Samachar Bhavan	India	1976	
SI	Sportinformation SI	Switzerland	1922	2016
SPK	Schweizerische Politische Korrespondenz	Switzerland	1917	1993
SPT	Schweizer Press-Telegraph	Swiss		
SPTA	Sankt-Petersburger Telegrafenagentur	Russia	1904	1918
Stefani	Agenzia Stefani	Italy	1853	1945
STT/FNB	Oy Suomen Tietotoimisto Finska Notisbyran AG	Finland	1887	
TANJUG	Telegrafska Agencija nova Jugoslavija	Yugoslavia/Serbia	1943	2018

Continuation of Table B.1 (Queried news agencies)

Abbr.	News Agency	Country	Creation	End
TASS	Telegrafnoie Agenstvo sovietskavo Soyusa	Russia	1925	
Telunion	Telegraphen-Union	Germany	1913	1934
TT	Tidningarnas Telegrambyra	Sweden	1921	
Ukrinform		Ukraine	1918	
UP-UPI	United Press (International)	USA	1907	1990
Wolffsbüro	Wolffs Telegraphisches Bureau	Germany	1849	1934
	Mixed Agency Content	Europe		
	Agence de presse Inter-France	France	1937	1944
	Agence Roumaine	Romania	1916	1921
	Agence Indo-Pacifique			
	Agence Balcanique	Bulgaria		
	Boesmanns Telegraphisches Bureau	Germany	1856	?
	Agence de Constantinople			
	Agence Fabre	Spain		
	Nordische Telegraphenbureau (Petersburg)	Russia	1869	
	Agence de Belgrad	Serbia		

TABLE B.2

Table containing the query details for the search of agencies in *impresso* and the storage of the agency collections, sorted by the amount of search hits per agency.

Begin of Table B.2 (Queried Agencies and Agency Collections)

News Agency Abbr.	#articles found in *impresso*	#articles stored	Collection Name	Query
ATS / SDA	1,051,964	1,033,440	ATS1, ATS2, ATS3	containing ats or atsj or atsl or atsi or fatsï or fatsl or atsf or agence telegraphique suisse or schweizerische depeschenagentur or sda
AFP	676,775	661,686	AFP1, AFP2	containing afp or afpj or afpl or fafp or iafp or agence france presse
AP	536,528	525,475	Associated Press, AP1, AP2	containing ap or associated press or vassociated or associated or yassociated or lassociated or dassociated or ass. press or assoc. press

Continuation of Table B.2 (Queried Agencies and Agency Collections)

News Agency Abbr.	#articles found in impresso	#articles stored	Collection Name	Query
REUTERS	503,333	477,373	Reuter, Reuters, reutersche	containing reutersche or reuterfche or neutersche or reuterschen or reuter'sche or neuterschen or reuterfchen or reuterbureau or reutermeldung or reuterbüro or reuters or reuterï or reutei or rcuter or rcutcr or reulers or reuteragentur or reutcr or reuterbüro or reutermeldung
HAVAS	217,013	217,013	Havas, Havasagentur	containing havas or bavas or hayas or havae or ilavas or llavas or flavas or haivas or uavas or havais or havasj or huvas or tlavas or jlavas or fhavas or lavas or havat or havaa or tiavas or haoas or hawas or haveis or liavas or havasl or hivas or hauas or iiavas or llaoas or heivas or havai or havasi or lfavas or mavas or hava or havasagentur or havasmeldung or haoasagentur or havasnote
UPI	161,447	153,403	UPI	containing upi or united press or united prefz
Wolffsbüro	95,416	93,215	Wolff	containing wolff or lwolff or lwolsf or lwolss or iwolff or wofff or wolffsbüro or wolffbüro or wolffbüros or wolffbureau or wolffbureaus or wolff'sche or wolffmeldung
DPA	77,060	75,173	DPA	containing dpa or deutsche presse agentur
TASS	76,061	75,006	TASS	containing tass or tafz or telegraphen-agentur or itar
DNB	44,979	44,155	DNB	containing nachrichtenbüro or nachrichtenbureau or richtenbüro or richtenbureau or dnb or ldnb or ldnv or dnib or idrb or d0lb or dnv0 or idnb or d0lv
Stefani	29,535	29,167	Stefani	containing stefani or steîani or stelani or stcfani or slefani or stefanl
SPK	25,821	25,701	SPK	containing spk or schweizerische politische korrespondenz or schweizer mittelpresse or smp

Continuation of Table B.2 (Queried Agencies and Agency Collections)

News Agency Abbr.	**#articles found in impresso**	**#articles stored**	**Collection Name**	**Query**
TT	12,582	12,579	TT-Sweden	containing tt AND suedois or suedoise or schwedisch or schwedische
PTI	12,285	12,208	PTI	containing pti or press trust of india
BELGA	11,139	10,954	Belga	containing belga or beiga
ANSA	9,949	9,882	ANSA	containing ansa or agenzia nationale
APA	8,871	8,799	APA	containing apa or austria press
Europapress	8,214	7,929	Europapress	containing europapreß or europapreh or europapretz or europaprefz or leuropapreß or eurovapreß or europapieß or eurovapreh or seuropapreß or leuropapreh or europapieh or europaprch or leuropapretz or euiopapreß or guropapreß
DDP/DAPD	6,671	6,633	DDP-DAPD	containing ddp or dapd or deutscher depeschendienst
	6,104	5,942	Mixed	containing afpreuters or afpreuter or atsafp or atsreuters or atsreuter or atsjafp or atsap or aplddp or aplafp or afplap or dpalafp or atsjreuter or atsfafp or ddplap or aplsda or aplddp or sdalafp or atsjréd or atsréd
TANJUG	5434	5361	Tanjug	containing tanjug or tanyoug or tanjoug
ANP	5,022	5,017	ANP	containing anp or algemeen nederland; not containing nationale - tagged as article
DOMEI	4,770	4,751	Domei	containing domei or domci
CTK	4,641	4,635	CTK	containing ctk or ceteka
Extel	4,186	4,184	Extel	containing extel or exlel
Telunion	4,186	4,143	Telegraphen-Union	containing telunion or ltelunion or stelunion or telunwn or telegraphen union
PAP	3,029	3,025	PAP	containing pap AND polonais or polnisch
Interfax	2,809	2,763	Interfax	containing interfax
LUSA	2,189			containing lusa - tagged as article (still too noisy)
BTA	2,158	2,156	BTA	containing bta or agence bulgare AND bulgare or bulgarisch

Continuation of Table B.2 (Queried Agencies and Agency Collections)

News Agency Abbr.	#articles found in impresso	#articles stored	Collection Name	Query
PA	1,748			containing press association
EPD	1,663			containing epd (very noisy)
Fides	1,523			containing fides AND agence or agentur
NTB	1,295			containing ntb - tagged as article
EFE	1,098			containing efe AND agence or espagnol or espagnole
Fournier	967			containing agence fournier or agence fourniei or agence foumier or lfournier or agence faurnier
AGERPRESS	908			containing agerpress or agerpres
NCNA/ Xinhua	670			
	522			containing inter-france
SPT	482			containing presstelegraph
	439			
ROSTA	422			containing rosta
CAPA	370			containing capa AND agence
Europa Press	308			containing europa press
INS	228			
Office-Correspondance	182			containing agence bullier or correspondance bullier or c. bullier or agence buliier or correspondance buliier
SPTA	149			containing spta or petersburger telegrafenagentur
JIJI	145			
SI	113			containing sportinformation si
ACP	102			containing acp and agence (only acp too noisy)
STT/FNB	94			containing stt or fnb AND finlandais or finnois or finlandaise or finlandai
ADN	92			containing adn and nachrichtendienst (adn too noisy)
CP	89			
Meurisse	79			containing meurisse published from Jan 1, 1880 to Dec 31, 1937
SAMACHAR	79			
Ottomane (?)	77			
KYODO	74			
RB / RITZAUS	71			containing ritzaus

Continuation of Table B.2 (Queried Agencies and Agency Collections)

News Agency Abbr.	**#articles found in impresso**	**#articles stored**	**Collection Name**	**Query**
	67			
Adnkronos	63			containing adnkronos
	29			containing boesmanns
	21			containing Agence de constantinople or mgence de constantinople
KKTK/ Korbureau	19			containing korrbureau or telegraphen-korrespondenz
Ukrinform	8			containing ukrinform
	2			containing nordische telegraphenbureau or telegraphenbureau petersburg
Rol	0			containing agence rol
BelTA	0			
MTI	0			

C Annotation Settings

C.1 Annotation Tagset

TABLE C.1
Tagset for the Annotation of the training corpus in Inception, with the tag description provided for the annotators

Begin of Table C.1 (Annotation Tagset)

	tag_name	tag_description
0	AFP	Agence France Presse, 1944-today (other variations: A.F.P.)
1	ANP	Algemeen Nederlands Persbureau, 1934-today
2	ANSA	Agenzia Nationale Stampa Associata, 1945-today
3	AP	Associated Press, 1848-today (other variations: Assoc. Press)
4	APA	Austria Press Agentur, 1946-today
5	ATS-SDA	Agence télégraphique suisse/Schweizerische Depeschenagentur, 1894-today (other variations: ATS, SDA)
6	BTA	Bulgarska Telegrafitscheka Agentzia, 1898-today (other variations: Agence Bulgare)
7	Belga	Agence Belga SA, 1920-today
8	CTK	Czechoslavenska Tiskova Kancelar, 1918-today (other variations: Ceteka, Agence Ceteka)
9	DDP-DAPD	Deutscher Depeschendienst, 1971-2009; Deutscher Auslands-Depeschendienst, 2009-2013 (other variations: DDP, DAPD)
10	DNB	Deutsches Nachrichtenbüro GmbH, 1933-1945 (other variations: D.N.B.)
11	DPA	Deutsche Presse Agentur, 1949-today
12	Domei	Domei Tsushin/Domein News Agency, 1936-1945 ((Japan))
13	Europapress	Europapress (other variations: Europapreß, Europapr.)
14	Extel	Exchange Telegraph Co. Ltd., 1872-1993 (other variations: Agence Extel)
15	Havas	Havas, 1835-today (other variations: Agence Havas)
16	Interfax	Interfax News Agency, 1989-today
17	PAP	Polska Agencja Prasowa, 1944-today
18	Reuters	Reuters, 1851-today (other variations: Reuter, Reutermeldung, Reuter'sche Bureau)
19	SPK-SMP	Schweizer Mittelpresse, 1917-1947; Schweizerische Politische Korrespondenz, 1947-1993 (other variations: SPK, SMP)
20	Stefani	Agenzia Stefani, 1853-1945 (other variations: Agence Stefani)
21	TANJUG	Telegrafska Agencija nova Jugoslavija, 1943-2018

Continuation of Table C.1 (Annotation Tagset)

	tag_name	tag_description
22	TASS	Telegrafnoie Agenstvo sovietskavo Soyusa, 1925-today (other variations: ITAR-TASS, Taß, Telegraphen-Agentur der Sowietunion, Telegraphen-Agentur der U.S.S.R.)
23	TT	Tidningarnas Telegrambyra (SE), 1921-today
24	Telunion	Telegraphen-Union, 1913, 1934 (other variations: TU)
25	UP-UPI	United Press, 1907-1958; United Press International, 1958-1990 (other variations: UP, UPI, United Press/Preß, United Press/Preß International)
26	Wolff	Wolffs Telegraphisches Bureau, 1849-1934 (other variations: Wolffagentur, Wolffsbüro, Kontinental-Telegraphencompagnie, Continental-Telegraphen-Compagnie Wolff's Telegraphisches Büro)
27	ag	for an agency mention without specifying the agency, i.e. ag. or agence, Agentur
28	pers.ind.articleauthor	author of the newspaper article
29	unk	unknown: any Agency whose name is not contained in Tagset

C.2 Annotation Guidelines

The following pages contain the annotation guidelines which were given out to the annotators during the annotation campaign.

News Agency Classification

Annotation Guidelines

April 2023

Lea Marxen

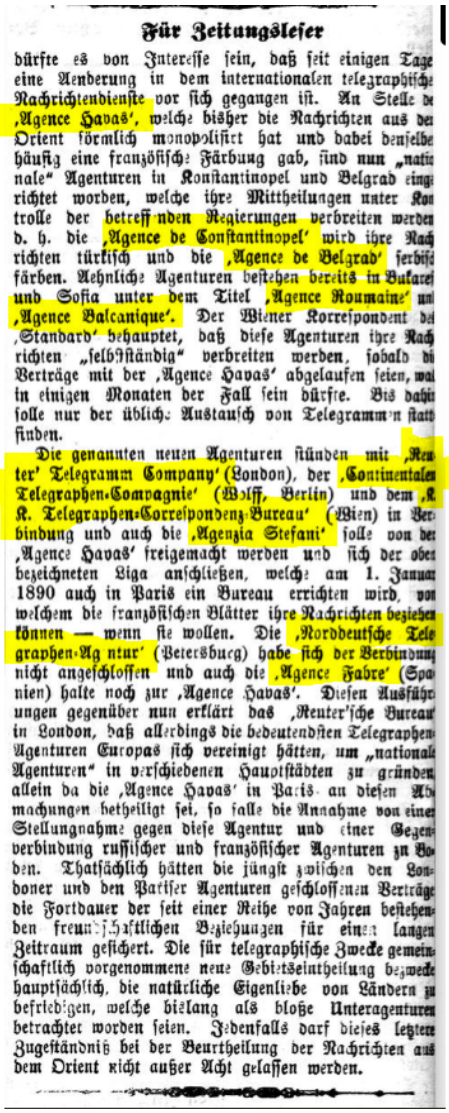
In the following, you will find some guidelines and comments on how to annotate the corpus for the *impresso* News Agency Classification.

1. News Agency Labels

The News Agency Labels can be found in the layer "*Impresso* News Agencies".

1.1 News Agencies

Only annotate News Agencies when they provide the information contained in the articles. News Agency mentions which occur because the content is *about* the News Agency but not *from* them should not be annotated, e.g. if the article treats a change of personnel within a News Agency or an acquisition of a News Agency by another (and this information does not come from the News Agency itself), see also Fig. 1.



Für Zeitungsleser

dürfte es von Interesse sein, daß seit einigen Tage eine Aenderung in dem internationalen telegraphische Nachrichtendienste vor sich gegangen ist. An Stelle de ‚Agence Havas', welche bisher die Nachrichten aus der Orient förmlich monopolisirt hat und dabei denselbe häufig eine französische Färbung gab, sind nun „natio nale" Agenturen in Konstantinopel und Belgrad eing richtet worden, welche ihre Mittheilungen unter Kon trolle der betreffenden Regierungen verbreiten werden d. h. die ‚Agence de Constantinopel' wird ihre Nach richten türkisch und die ‚Agence de Belgrad' serbisc färben. Aehnliche Agenturen bestehen bereits in Bukarest und Sofia unter dem Titel ‚Agence Roumaine' un ‚Agence Balcanique'. Der Wiener Korrespondent des ‚Standard' behauptet, daß diese Agenturen ihre Nach richten „selbstständig" verbreiten werden, sobald di Verträge mit der ‚Agence Havas' abgelaufen seien, wa in einigen Monaten der Fall sein dürfte. Bis dahin solle nur der übliche Austausch von Telegrammen statt finden.

Die genannten neuen Agenturen stünden mit ‚Reu ter' Telegramm Company' (London), der ‚Continentalen Telegraphen-Compagnie' (Wolff, Berlin) und dem ‚K. K. Telegraphen-Correspondenz-Bureau' (Wien) in Ver bindung und auch die ‚Agenzia Stefani' solle von der ‚Agence Havas' freigemacht werden und sich der oben bezeichneten Liga anschließen, welche am 1. Januar 1890 auch in Paris ein Bureau errichten wird, von welchem die französischen Blätter ihre Nachrichten beziehen können — wenn sie wollen. Die ‚Norddeutsche Tele graphen-Agentur' (Petersburg) habe sich der Verbindung nicht angeschlossen und auch die ‚Agence Fabre' (Spa nien) halte noch zur ‚Agence Havas'. Diesen Ausführ ungen gegenüber nun erklärt das ‚Reuter'sche Bureau' in London, daß allerdings die bedeutendsten Telegraphen Agenturen Europas sich vereinigt hätten, um „nationale Agenturen" in verschiedenen Hauptstädten zu gründen, allein da die ‚Agence Havas' in Paris an diesen Ab machungen betheiligt sei, so falle die Annahme von einer Stellungnahme gegen diese Agentur und einer Gegen verbindung russischer und französischer Agenturen zu Bo den. Thatsächlich hätten die jüngst zwischen den Lon doner und den Pariser Agenturen geschlossenen Verträge die Fortdauer der seit einer Reihe von Jahren bestehen den freundschaftlichen Beziehungen für einen langen Zeitraum gesichert. Die für telegraphische Zwecke gemein schaftlich vorgenommene neue Gebietseintheilung bezwecke hauptsächlich, die natürliche Eigenliebe von Ländern zu befriedigen, welche bislang als bloße Unteragenturen betrachtet worden seien. Jedenfalls darf dieses letzteren Zugeständniß bei der Beurtheilung der Nachrichten aus dem Orient nicht außer Acht gelassen werden.

Fig. 1: Example for News Agency mentions which should not be annotated.

C.2 Annotation Guidelines

The following News Agency labels exist for annotation:

Abbr.	Other possible Reference In Newspapers	News Agency	Country	Creation	End
AFP	(AFP), (Afp.), (A. F. P.), Agence France Presse	Agence France-Presse	France	1944	
ANP		Algemeen Nederlands Persbureau	Netherlands	1934	
ANSA		Agenzia Nationale Stampa Associata	Italy	1945	
AP	(AP)	Associated Press	USA	1848	
APA		Austria Press Agentur	Austria	1946	
ATS / SDA	Agence télégraphique suisse	Agence Telegraphique Suisse / Schweizerische Depeschenagentur	Switzerland	1894	
BELGA	(Belga)	Agence télégraphique belge de Presse	Belgium	1920	
BTA	Agence Bulgare, BTA	Bulgarska Telegrafitscheka Agentzia / Agence Bulgare / Bulgarian Telegraph Agency	Bulgaria	1898	
CTK		Czechoslavenska Tiskova Kancelar / Agence Ceteka	Czechoslovakia/Czech	1918	
DDP / DAPD	Ddp, dapd	Deutscher Depeschendienst / Deutscher Auslands-Depeschendienst	Germany	1971	2013
DNB	(D. N. B.)	Deutsches Nachrichtenbüro GmbH	Germany	1933	1945
DOMEI	(Domei.)	Domei Tsushin/Domei News Agency	Japan	1936	1945
DPA	(DPA)	Deutsche Presse Agentur	Germany (GFR)	1949	
Europapress	(Europapreß.), (Europapr.)	Europapress	Germany		
Extel	Agence Extel, (Extel.)	Exchange Telegraph Co. Ltd.	United Kingdom	1872	1993
HAVAS	(Havas.), Havas	Havas	France	1835	
Interfax		Interfax News Agency	Russia	1989	
PAP		Polska Agencja Prasowa	Poland	1944	
REUTERS	(Reuter), (Reuter.), Reutermeldung, Reuter'schen Büreau, Reuter-Telegramm	Reuters	United Kingdom	1851	

SPK / SMP		Schweizerische Politische Korrespondenz / Schweizer Mittelpresse (until 1947)	Switzerland	1917	1993
Stefani	Agenzia Stefani, (Stefani.)	Agenzia Stefani	Italy	1853	1945
TANJUG		Telegrafska Agencija nova Jugoslavija	Yugoslavia/Serbia	1943	2018
TASS	ITAR-TASS, Taß, Telegraphen-Agentur der Sowietunion, Telegraphen-Agentur der U.S.S.R.	Telegrafnoie Agenstvo sovietskavo Soyusa	Russia	1925	
TT		Tidningarnas Telegrambyra	Sweden	1921	
Telunion	TU, (Telunion.)	Telegraphen-Union	Germany	1913	1934
UPI		United Press International	USA	1958	1990
Wolff	(Wolff.), Wolffbüro, Wolffsbüro, Kontinental-Telegraphencompagnie	Wolffs Telegraphisches Bureau	Germany	1849	1934

1.2 The *unk* label

The *unk* label should be used for News Agencies whose name is not contained in the predefined labels. If this is the case, and you are sure that you are tagging a news agency, report the annotation as successful when finishing it. If you are unsure whether you are tagging a news agency or an article author, please report it as a problem when finishing the annotation of the document and type “unk” in the message field:

Fig. 2: Press the indicated button to finish the annotation of a document.

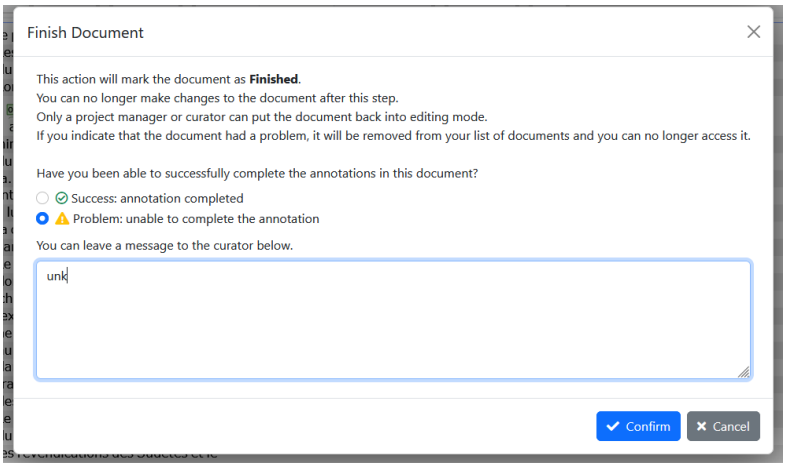


Fig. 3: Reporting a problem with message "unk" when finishing the document annotation

1.3 The *pers.ind.articleauthor* label

At some time, newspapers began crediting the authors of the articles. This can take the same form as crediting a News Agency, as can be seen in Figure 4. If you see this, use the *pers.ind.articleauthor* label to annotate the author of the article.

ÉTATS-UNIS

Grâce refusée à Troy Davis

La justice américaine a refusé mardi d'accorder sa grâce à Troy Davis, un Noir condamné à mort en 1991 pour le meurtre d'un policier blanc et devenu un symbole de la lutte contre la peine de mort. Cette décision tombe à la veille de son exécution prévue dans l'Etat de Géorgie. «*Le comité a refusé sa clémence*», a indiqué dans un communiqué le comité des grâces de Géorgie. La réunion de ce comité à Atlanta, la capitale de Géorgie, était considérée comme la dernière chance pour le condamné de voir sa peine de mort commuée en prison à vie, le gouverneur de l'Etat ne disposant pas du droit de grâce. L'exécution de Troy Davis par injection mortelle est programmée jeudi à 1h en Suisse à la prison de Jackson, malgré des doutes sur sa culpabilité. ATS-REUTERS

FRANCE

Vente de manuscrits de Gainsbourg

Les manuscrits des chansons «Sorry Angel», «Love on the beat», «No Comment», «Hmm Hmm Hmm» et de «You're under arrest» seront mis en vente par Sotheby's Paris le 9 novembre prochain. Des notes diverses, des tapuscrits ainsi que des photos inédites de Serge Gainsbourg (1928-1991) réalisées pour un magazine anglais compléteront cette vente. Le brouillon manuscrit de «Love on the beat» (1984) est peut-être le plus fascinant. Selon les organisateurs de la vente, ces deux pages comportent de très nombreuses variantes et corrections faisant apparaître en filigrane les influences baudelairiennes présentes dans toute l'œuvre de l'artiste (estimation: entre 12 000 et 18 000 euros). PHV

NIDWALD

Un soldat blessé lors d'un exercice

La justice militaire enquête depuis un mois sur un incendie qui a fait un blessé lors d'un exercice à Oberdorf (NW), près de Stans. Souffrant de blessures légères à moyennes, ce soldat en cours de répétition avait reçu l'ordre d'allumer un feu destiné à être éteint par des soldats au centre de formation de la Swisscoy. Les faits se sont produits le 23 août dernier, a indiqué hier la justice militaire. Le blessé n'était pas membre de la troupe qui devait être envoyée au Kosovo. Motocycliste, il faisait partie d'un détachement responsable de l'exploitation des installations du centre de formation Swissint. Les circonstances de l'accident font l'objet d'une enquête. L'investigation en est à son stade initial. Des spécialistes de l'institut médico-légal de la ville de Zurich se sont joints aux enquêteurs. ATS

Fig. 4: Articles with News Agency credentials (yellow) and author credentials (red)

If you are unsure whether a reference belongs to a News Agency or the article other, label the token with *unk* and follow the procedure from 1.2.

2. OCR noise

There exist two different options to deal with OCR noise, one on document level to discard the article and one on token level to correct the spelling of a News Agency mention.

2.1 Noisy article

If a document is too noisy to annotate, you can indicate this by setting the label *non_usable* to “Yes” in the Document Metadata. For this, go to “Document Metadata” in the menu bar on the left, as indicated in Figure 5.

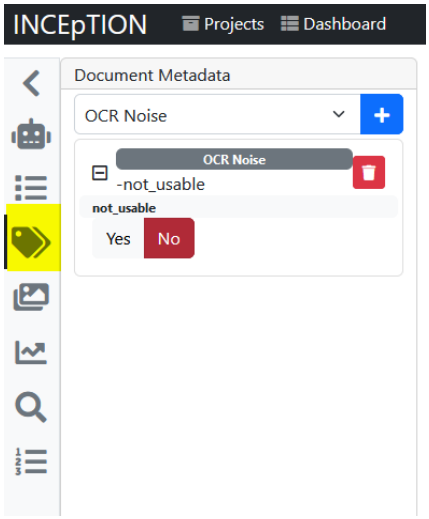


Fig. 5: Location of Document Metadata

2.2 Noisy News Agency Mention

If the OCR transcription of a News Agency mention is incorrect, please correct the transcription (and set the *noisy_ocr* tag to “Yes”).

Concerning the token boundaries, we will follow the specifications from the “Named Entity Annotation Guidelines” (HIPE):

Special cases with noisy OCR:
When it is difficult to establish the boundary of a mention because of noisy OCR:

- *look at the image*
- *include, in the annotation, the garbage characters which you think should have been recognized and should be part of the mention*

- *mark the mention with the flag “noisy-entity” and add your OCR hypothesis correction.*

*ex: in the string Trève * (which stands for Trèves), the full string Trève * should be annotated, not only Trève.*

3. Token boundaries

This section clarifies which part of a News Agency mention should be marked for annotation.

3.1 Abbreviations

Specification when to include "." in annotation:

- **Include**: abbreviation of a name (e.g. "ag.", "Ag. Télégr. Suisse", "D.N.B.")
- **Do not include:** at the end of an agency name (e.g. Havas. → Havas) or the acronym without points in between (e.g. AFP. → AFP)

3.2 Compounds / Proper Names

Do not include the words "agence" or "Agentur" in the annotation (e.g. agence [**Extel**]), except if they belong to a proper name (e.g. Agence France Presse, Agence Télégraphique Suisse).

Regarding German compounds, only include the name of the agency, e.g. [**Reuter**]meldung, [**Reuter**]-Telegramm.
Also note the following instructions from the "Named Entity Annotation Guidelines" (HIPE):

Special case with German compounds: Apply the cross-lingual or decomposition test, i.e. translate the compound to French and in the German compound annotate only what should be annotated in French.

The connecting "s" in German compounds is not annotated:
Völkerbundsmitgliedern
=> only Völkerbund is annotated
< org .adm> Völkerbund </ org .adm> smitgliedern

4. Remaining Annotation Issues

If it remains unclear how to annotate a certain mention, note the issue in the following file:

Google Document for Annotation issues

Resolving issues jointly and thus creating further examples of “best practice" can also ensure that the annotations will be as consistent as possible.

D Dataset Visualizations (Additional Material)

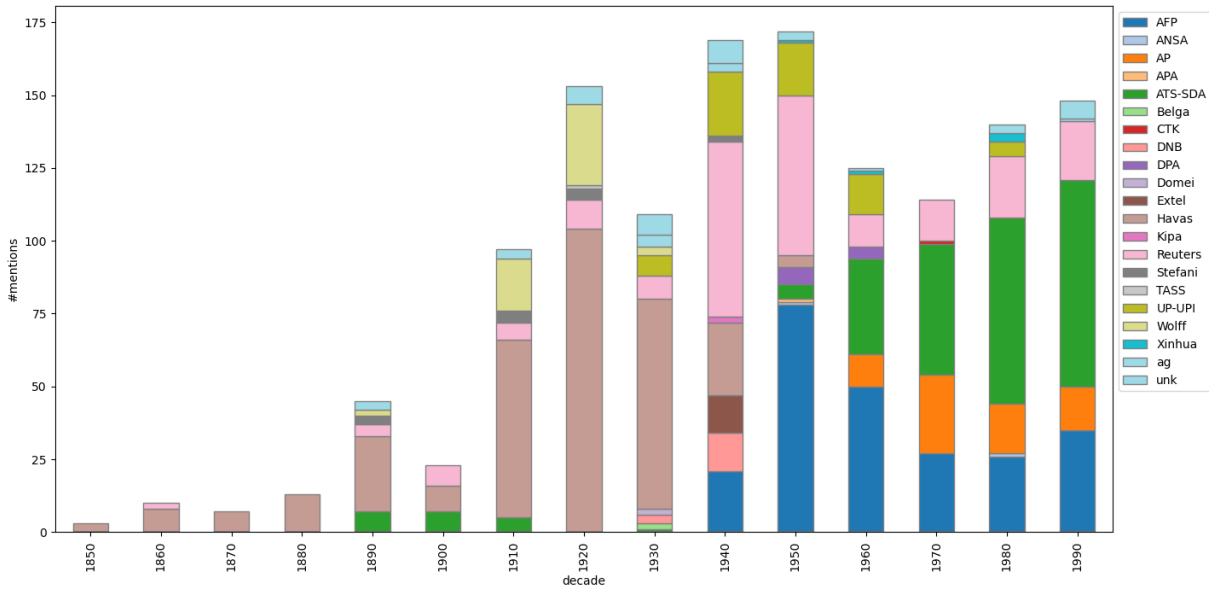


FIGURE D.1
Occurrence of different news agency mentions in the French annotated dataset over time.

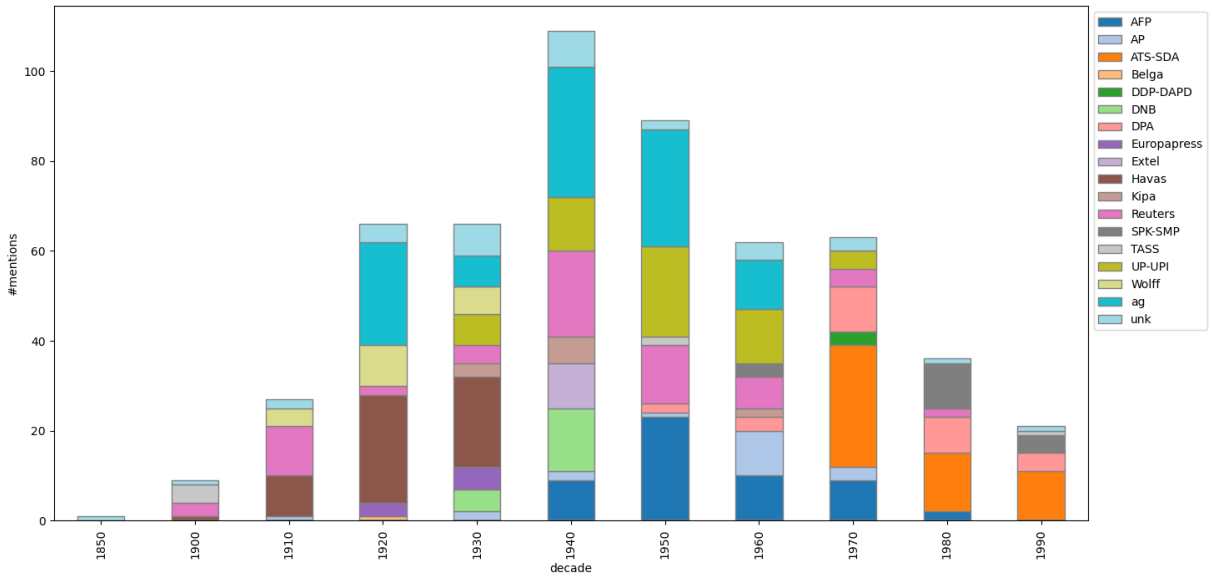


FIGURE D.2
Occurrence of different news agency mentions in the German annotated dataset over time.

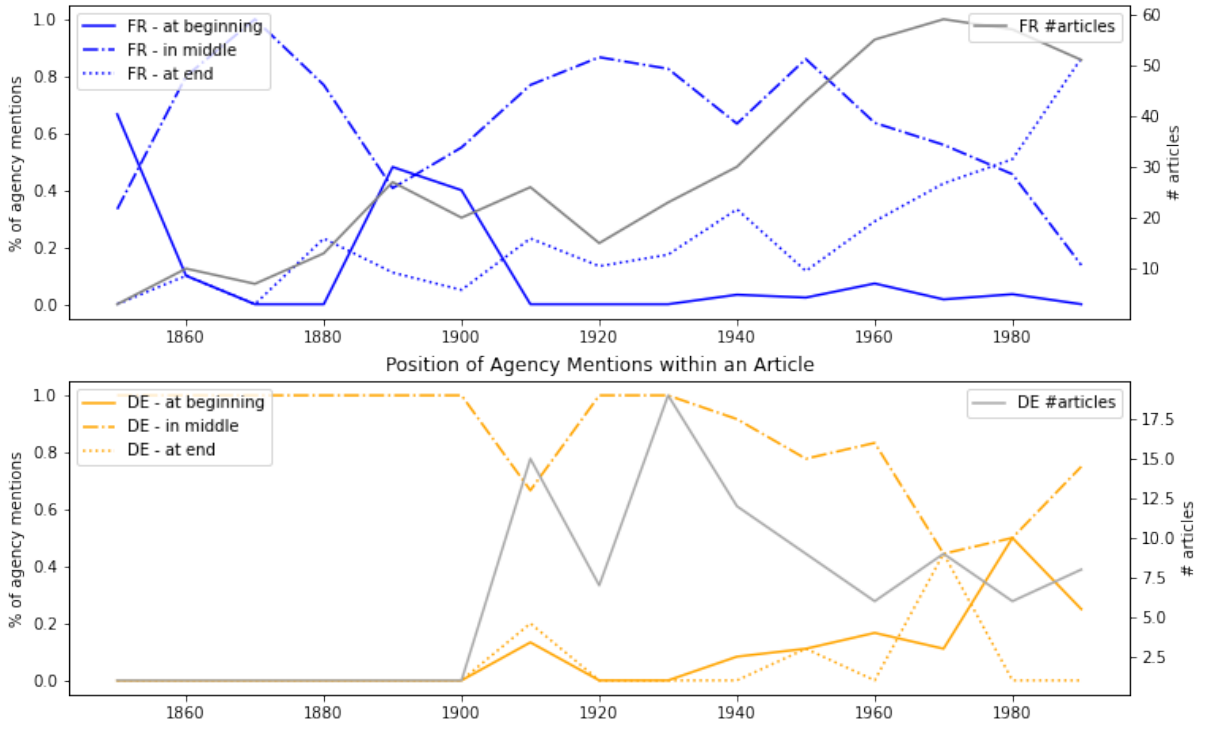


FIGURE D.3
Positions of agency mentions within an article in the annotated dataset over time, split by language (French above, German below). The grey line shows the total number of articles per decade (right y-axis). Agency mentions in the middle are probably over-represented because the OLR often combines short consecutive articles into one big article.

E Experimental Results

Overview of Figures and Tables in this Section, in the order of their appearance:

E.1 In-model vs. HIPE Evaluation

For the task of NER, we disposed of two evaluation scorers, the in-model evaluation scorer which took the input of the model as a basis, and the HIPE-scorer from the HIPE campaign[1]. We expected the HIPE-scorer to give slightly worse evaluations, as it also incorporated those agency mentions that were discarded by the model due to its dependence on the maximum sequence length. However, the opposite was the case: The average over all models of the difference $F1_{\text{HIPE}} - F1_{\text{in-model}}$ amounted to 0.020 for the French models and 0.028 for the German models. The minimum of the difference lay at -0.041 for French and -0.036 for German, thus indicating that the HIPE evaluation did not always yield higher results, but this was not the majority, as the median values of 0.020 for French and 0.018 for German show. The comparison across maximum sequence lengths in Table E.1 reveals that the HIPE scorer seems to provide higher performance results with growing sequence length. Only the the average over the German models for the maximum sequence length of 256 diverges from this trend.

TABLE E.1
Difference $F1_{\text{HIPE}} - F1_{\text{in-model}}$, averaged over all models for different languages and maximum sequence lengths respectively.

Language	**Maximum Sequence Length**			
	64	**128**	**256**	**512**
de	0.019	0.032	0.020	0.042
fr	0.002	0.022	0.025	0.031

[1] https://github.com/hipe-eval/HIPE-scorer

E.2 Named Entity Agency Recognition (Additional Material)

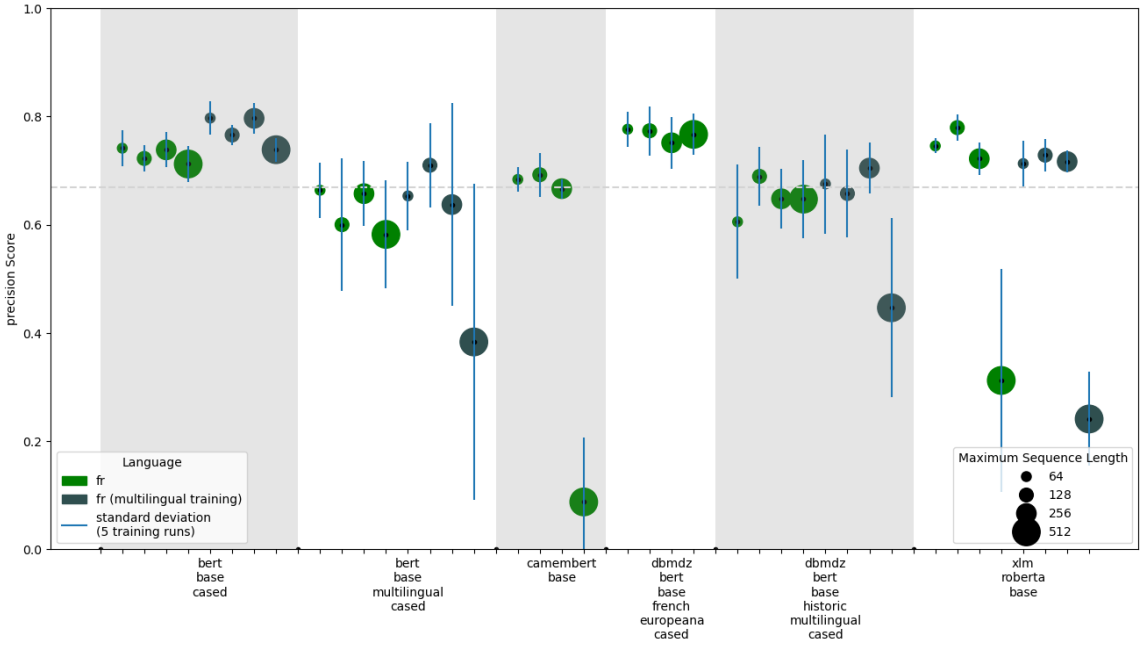


FIGURE E.1

Precision results for agency recognition on the French test set. Experiments were run five times per configuration, the dots present the mean, the blue lines the standard deviation. The colour of the dots refers to the training set (French or Multilingual), while the size specifies the maximum sequence length. The lookup baseline is displayed as a grey dashed line (see Section 4.1.3).

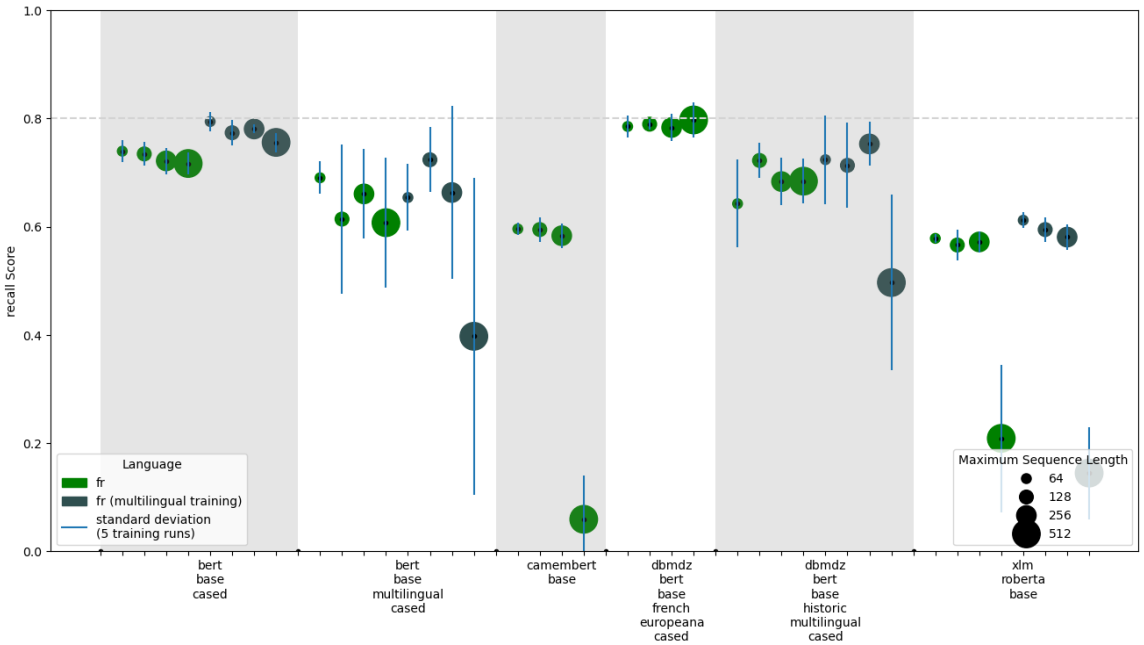


FIGURE E.2

Recall results for agency recognition on the French test set. The layout specifications are the same as for Figure E.1.

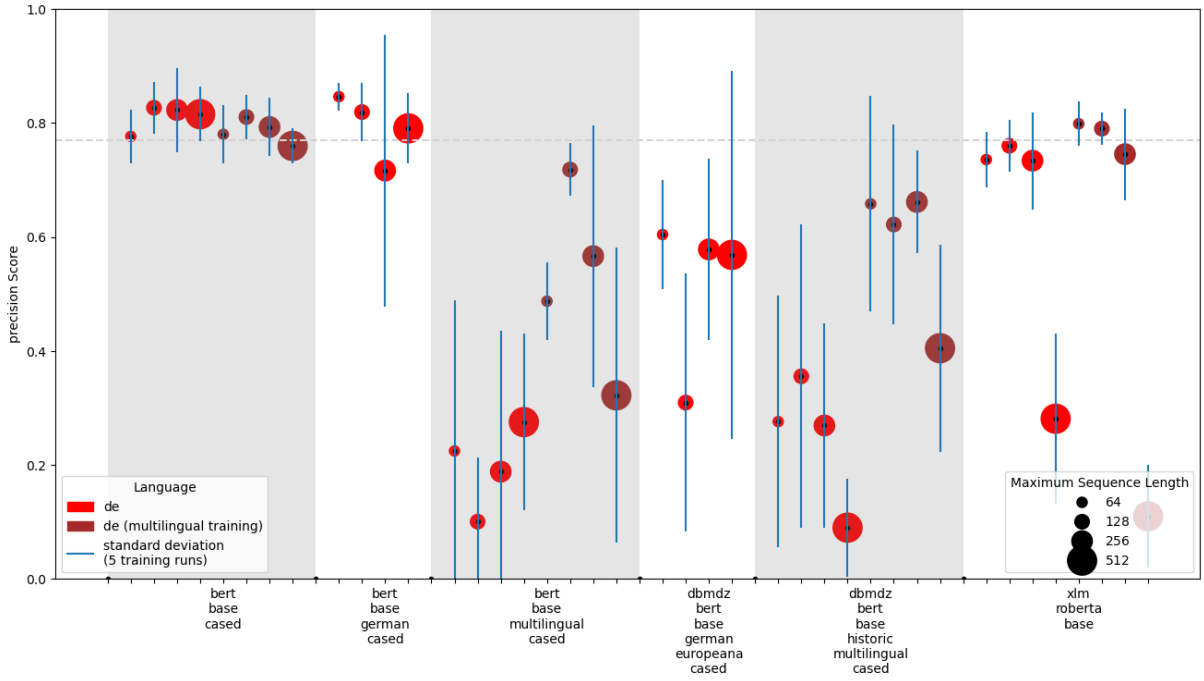


FIGURE E.3

Precision results for agency recognition on the German test set. Experiments were run five times per configuration, the dots present the mean, the blue lines the standard deviation. The colour of the dots refers to the training set (German or Multilingual), while the size specifies the maximum sequence length. The lookup baseline is displayed as a grey dashed line (see Section 4.1.3).

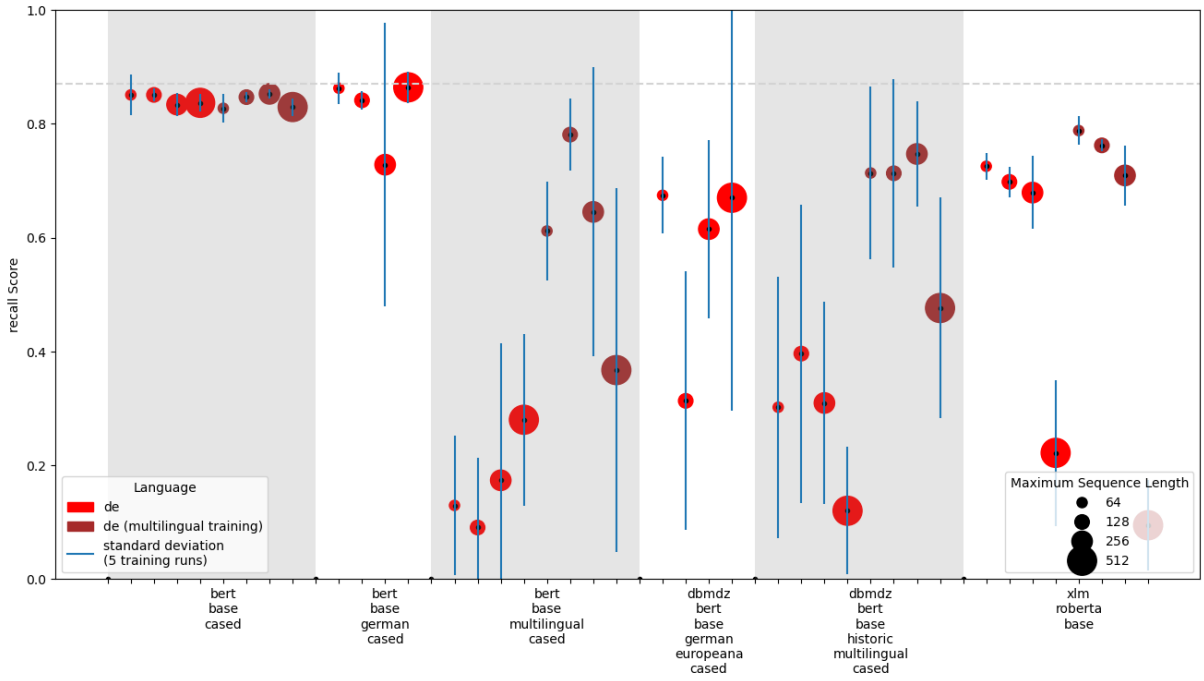


FIGURE E.4

Recall results for agency recognition on the German test set. The layout specifications are the same as for Figure E.3.

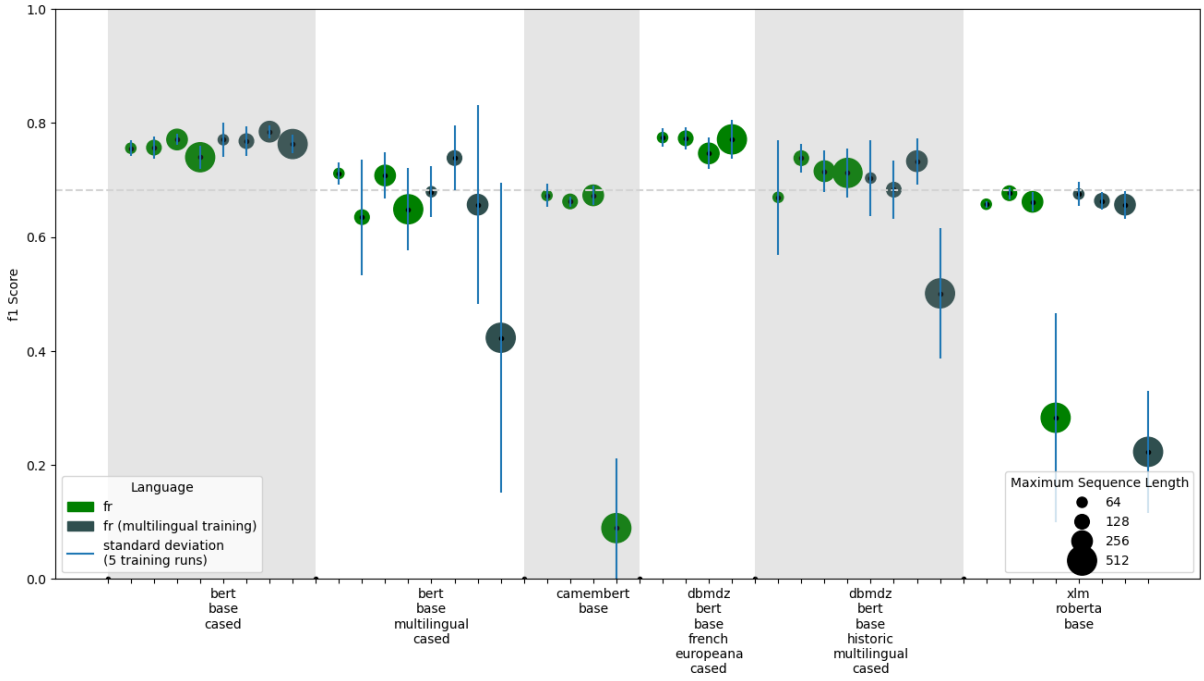


FIGURE E.5

F-scores for agency recognition on the French **dev** set. Experiments were run five times per configuration, the dots present the mean, the blue lines the standard deviation. The colour of the dots refers to the training set (French or Multilingual), while the size specifies the maximum sequence length. The lookup baseline is displayed as a grey dashed line (see Section 4.1.3).

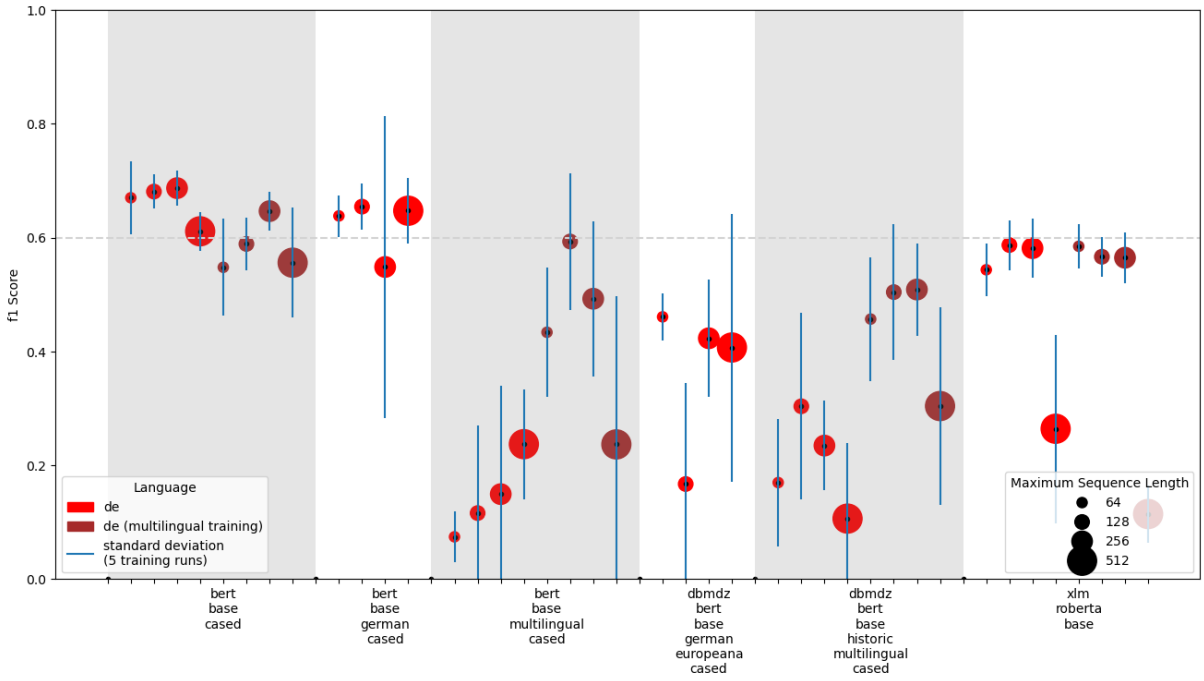


FIGURE E.6

F-scores for agency recognition on the German **dev** set. The layout specifications are the same as for Figure E.5.

TABLE E.2
Results for named entity (agency) recognition for the French test set. Experiments were run five times per configuration, the values show the mean and the standard deviation in brackets.

Language (train-eval)	Model	Max. Seq. Length	F1	Precision	Recall
fr-fr	bert_base_cased	64	0.740 (0.024)	0.742 (0.033)	0.740 (0.020)
		128	0.728 (0.022)	0.723 (0.025)	0.735 (0.022)
		256	0.730 (0.025)	0.738 (0.033)	0.722 (0.024)
		512	0.715 (0.025)	0.712 (0.034)	0.717 (0.020)
	bert_base_multilingual_cased	64	0.676 (0.038)	0.664 (0.051)	0.691 (0.030)
		128	0.607 (0.130)	0.600 (0.123)	0.614 (0.137)
		256	0.659 (0.071)	0.658 (0.060)	0.661 (0.083)
		512	0.594 (0.108)	0.582 (0.100)	0.608 (0.121)
	camembert_base	64	0.637 (0.010)	0.684 (0.023)	0.596 (0.011)
		128	0.640 (0.025)	0.692 (0.040)	0.595 (0.023)
		256	0.622 (0.019)	0.667 (0.018)	0.583 (0.023)
		512	0.071 (0.097)	0.087 (0.120)	0.059 (0.081)
	dbmdz_bert_base_french_europeana_cased	64	0.781 (0.025)	0.777 (0.032)	0.785 (0.020)
		128	0.781 (0.025)	0.774 (0.046)	0.790 (0.010)
		256	0.767 (0.036)	0.752 (0.048)	0.784 (0.025)
		512	0.782 (0.034)	0.767 (0.038)	0.797 (0.032)
	dbmdz_bert_base_historic_multilingual_cased	64	0.623 (0.094)	0.605 (0.105)	0.643 (0.081)
		128	0.705 (0.043)	0.689 (0.054)	0.723 (0.033)
		256	0.665 (0.048)	0.648 (0.055)	0.684 (0.044)
		512	0.665 (0.057)	0.647 (0.073)	0.684 (0.042)
	xlm_roberta_base	64	0.652 (0.009)	0.746 (0.014)	0.579 (0.009)
		128	0.655 (0.018)	0.780 (0.024)	0.566 (0.028)
		256	0.638 (0.023)	0.722 (0.030)	0.572 (0.019)
		512	0.250 (0.165)	0.312 (0.207)	0.209 (0.137)
multi-fr	bert_base_cased	64	0.796 (0.023)	0.797 (0.030)	0.795 (0.017)
		128	0.770 (0.019)	0.766 (0.018)	0.774 (0.023)
		256	0.789 (0.017)	0.797 (0.028)	0.781 (0.008)
		512	0.747 (0.015)	0.739 (0.022)	0.756 (0.018)
	bert_base_multilingual_cased	64	0.654 (0.062)	0.653 (0.063)	0.654 (0.062)
		128	0.717 (0.069)	0.710 (0.078)	0.724 (0.060)
		256	0.650 (0.175)	0.637 (0.187)	0.663 (0.160)
		512	0.390 (0.292)	0.383 (0.293)	0.397 (0.292)
	dbmdz_bert_base_historic_multilingual_cased	64	0.699 (0.087)	0.675 (0.092)	0.724 (0.082)
		128	0.684 (0.080)	0.658 (0.081)	0.714 (0.079)
		256	0.728 (0.043)	0.705 (0.047)	0.753 (0.041)
		512	0.470 (0.164)	0.446 (0.166)	0.497 (0.162)
	xlm_roberta_base	64	0.658 (0.020)	0.713 (0.042)	0.612 (0.015)
		128	0.655 (0.023)	0.729 (0.031)	0.595 (0.023)
		256	0.642 (0.022)	0.717 (0.020)	0.581 (0.024)
		512	0.177 (0.096)	0.241 (0.087)	0.145 (0.085)

TABLE E.3
Results for named entity (agency) recognition for the German test set. Experiments were run five times per configuration, the values show the mean and the standard deviation in brackets.

Language (train-eval)	Model	Max. Seq. Length	F1	Precision	Recall
de-de	bert_base_cased	64	0.812 (0.036)	0.777 (0.048)	0.851 (0.036)
		128	0.838 (0.027)	0.827 (0.045)	0.851 (0.011)
		256	0.828 (0.047)	0.823 (0.074)	0.834 (0.021)
		512	0.825 (0.022)	0.816 (0.048)	0.837 (0.015)
	bert_base_german_cased	64	0.854 (0.022)	0.846 (0.024)	0.863 (0.028)
		128	0.830 (0.029)	0.819 (0.051)	0.842 (0.017)
		256	0.722 (0.242)	0.717 (0.238)	0.728 (0.249)
		512	0.825 (0.046)	0.791 (0.062)	0.864 (0.027)
	bert_base_multilingual_cased	64	0.142 (0.124)	0.225 (0.265)	0.129 (0.122)
		128	0.093 (0.119)	0.100 (0.112)	0.091 (0.123)
		256	0.178 (0.245)	0.188 (0.247)	0.174 (0.241)
		512	0.277 (0.153)	0.275 (0.155)	0.280 (0.151)
	dbmdz_bert_base_german_europeana_cased	64	0.637 (0.081)	0.604 (0.096)	0.675 (0.067)
		128	0.311 (0.227)	0.309 (0.227)	0.313 (0.227)
		256	0.593 (0.150)	0.578 (0.159)	0.615 (0.157)
		512	0.615 (0.346)	0.569 (0.323)	0.670 (0.375)
	dbmdz_bert_base_historic_multilingual_cased	64	0.287 (0.225)	0.276 (0.221)	0.302 (0.229)
		128	0.374 (0.264)	0.356 (0.266)	0.396 (0.262)
		256	0.287 (0.179)	0.269 (0.180)	0.309 (0.178)
		512	0.103 (0.097)	0.090 (0.086)	0.120 (0.112)
	xlm_roberta_base	64	0.730 (0.033)	0.736 (0.049)	0.725 (0.024)
		128	0.727 (0.026)	0.760 (0.045)	0.698 (0.027)
		256	0.705 (0.071)	0.734 (0.085)	0.679 (0.064)
		512	0.247 (0.137)	0.281 (0.149)	0.222 (0.128)
multi-de	bert_base_cased	64	0.803 (0.036)	0.780 (0.051)	0.827 (0.026)
		128	0.828 (0.018)	0.811 (0.039)	0.847 (0.009)
		256	0.821 (0.025)	0.793 (0.051)	0.853 (0.008)
		512	0.793 (0.024)	0.760 (0.031)	0.830 (0.015)
	bert_base_multilingual_cased	64	0.543 (0.076)	0.487 (0.068)	0.612 (0.087)
		128	0.748 (0.053)	0.718 (0.046)	0.781 (0.063)
		256	0.603 (0.241)	0.567 (0.230)	0.645 (0.254)
		512	0.342 (0.286)	0.322 (0.259)	0.367 (0.320)
	dbmdz_bert_base_historic_multilingual_cased	64	0.684 (0.173)	0.658 (0.189)	0.714 (0.152)
		128	0.664 (0.172)	0.622 (0.175)	0.713 (0.165)
		256	0.702 (0.091)	0.662 (0.090)	0.747 (0.092)
		512	0.435 (0.185)	0.405 (0.181)	0.476 (0.194)
	xlm_roberta_base	64	0.793 (0.027)	0.799 (0.039)	0.788 (0.026)
		128	0.776 (0.009)	0.790 (0.029)	0.762 (0.010)
		256	0.727 (0.065)	0.745 (0.080)	0.709 (0.053)
		512	0.102 (0.084)	0.110 (0.090)	0.095 (0.080)

E.3 Sentence Classification (Additional Material)

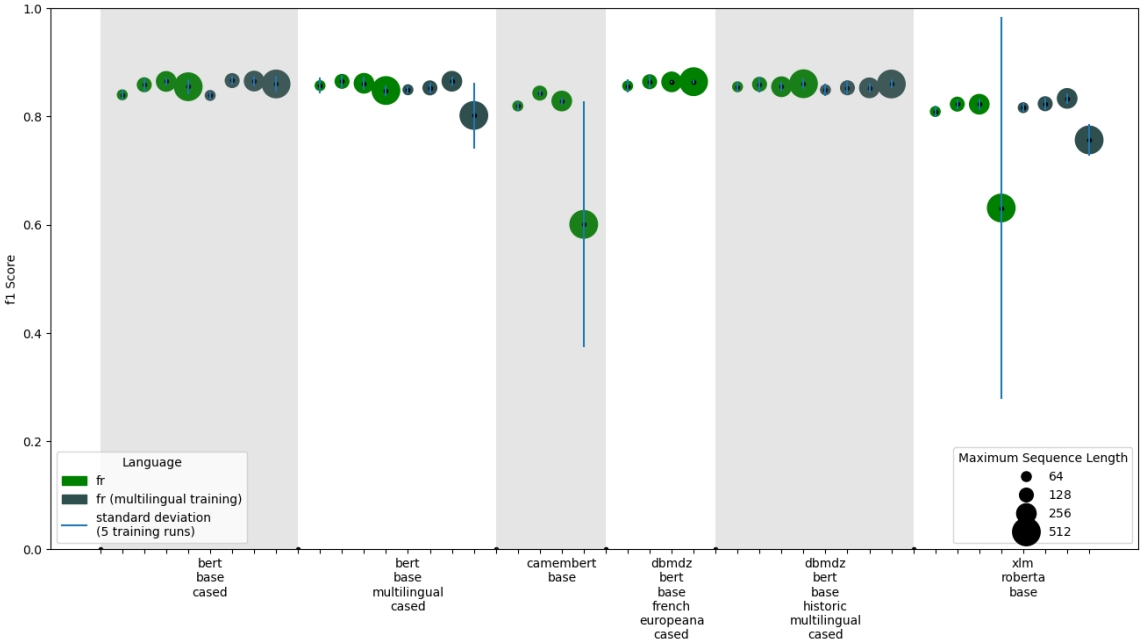


FIGURE E.7
Results for sentence classification on the French test set, showing the F-score for the "Positive" class, i.e. all sentences with an agency mention. Experiments were run five times per configuration, the dots present the mean, the blue lines the standard deviation. The colour of the dots refers to the training set (French or Multilingual), while the size specifies the maximum sequence length.

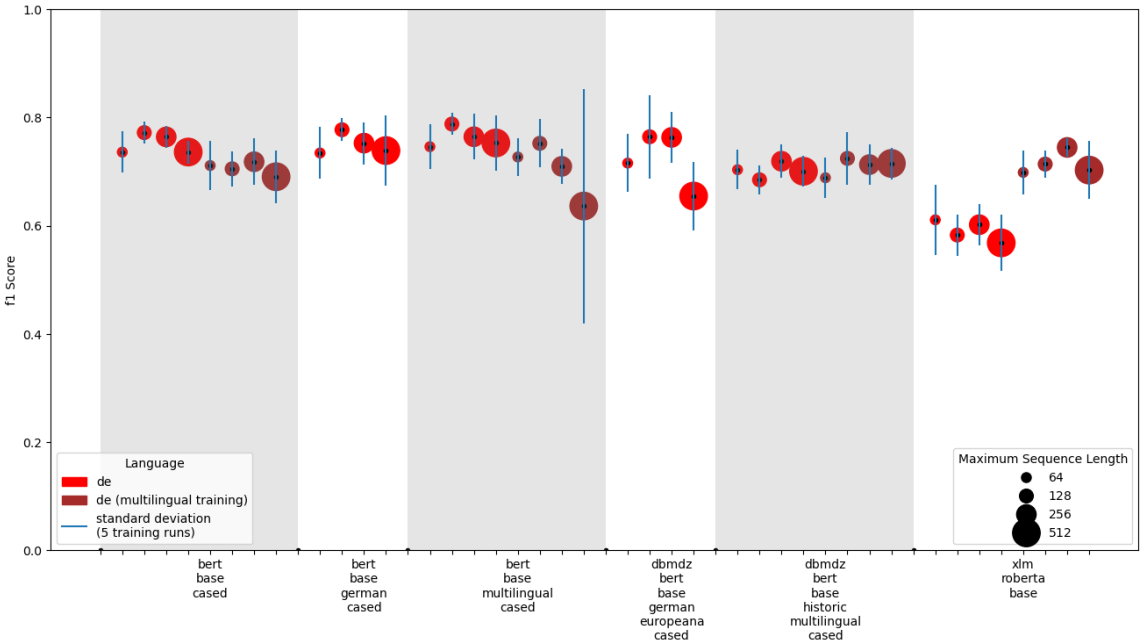


FIGURE E.8
Results for sentence classification on the German test set, showing the F-score for the "Positive" class, i.e. all sentences with an agency mention. the layout specifications are the same as for Figure E.7.

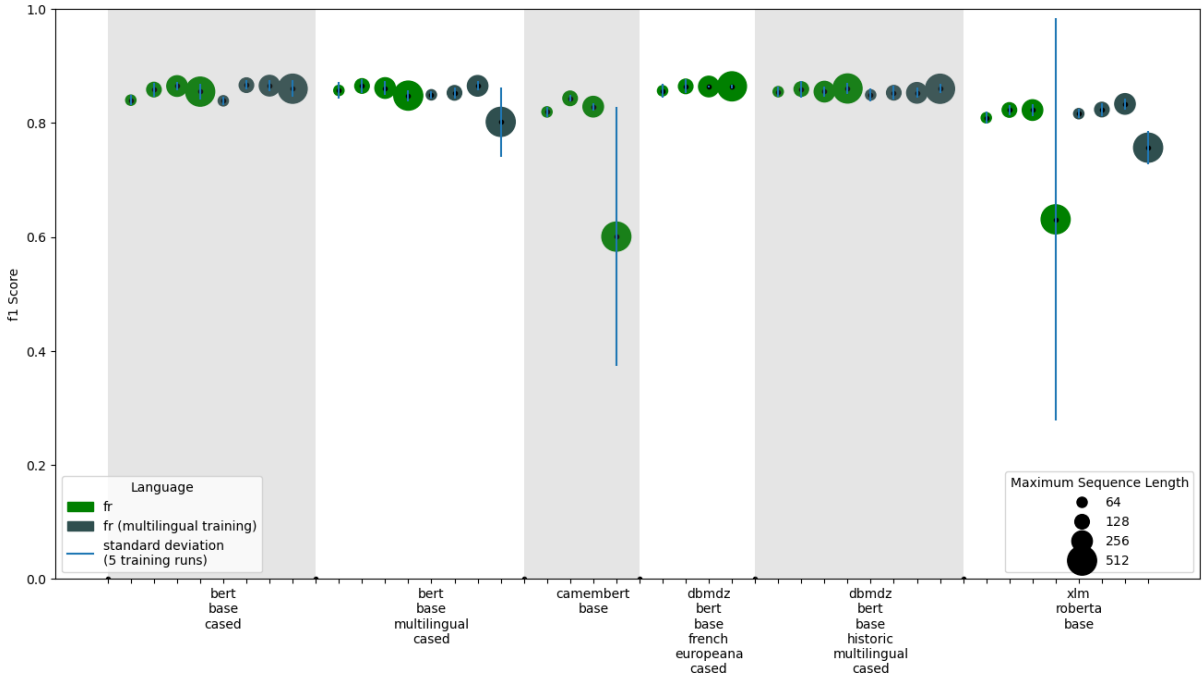


FIGURE E.9

Results for sentence classification on the French **dev** set, showing the F-score for the "Positive" class, i.e. all sentences with an agency mention. Experiments were run five times per configuration, the dots present the mean, the blue lines the standard deviation. The colour of the dots refers to the training set (French or Multilingual), while the size specifies the maximum sequence length.

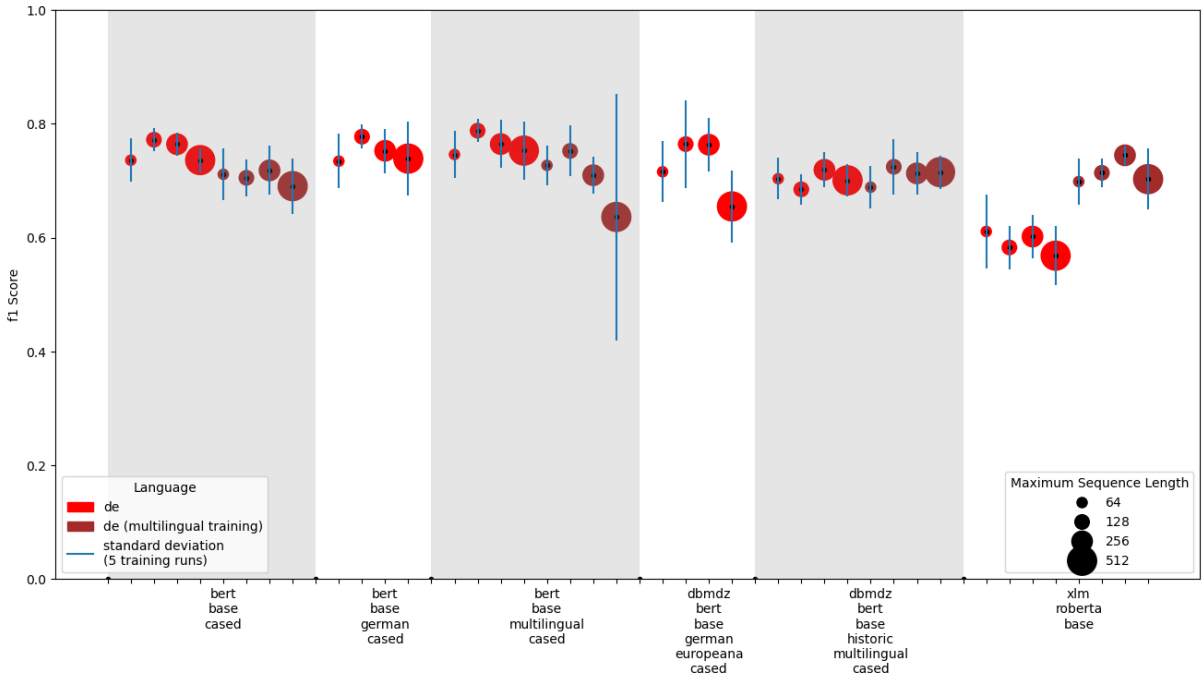


FIGURE E.10

Results for sentence classification on the German **dev** set, showing the F-score for the "Positive" class, i.e. all sentences with an agency mention. the layout specifications are the same as for Figure E.9.

TABLE E.4

Sentence classification results for the French test set. F-score both for the overall classification and for the class with the "Positives", i.e. those sentences with an agency mention. Experiments were run five times per configuration, the values show the mean and the standard deviation in brackets.

Language (train-eval)	Model	Max. Seq. Length	F1 (Both Classes)	F1 ("Positive" Class)
fr-fr	bert_base_cased	64	0.986 (0.001)	0.861 (0.014)
		128	0.988 (0.000)	0.877 (0.005)
		256	0.988 (0.001)	0.883 (0.012)
		512	0.986 (0.002)	0.868 (0.016)
	bert_base_multilingual_cased	64	0.989 (0.000)	0.885 (0.003)
		128	0.989 (0.001)	0.895 (0.010)
		256	0.989 (0.001)	0.895 (0.009)
		512	0.988 (0.001)	0.886 (0.006)
	camembert_base	64	0.983 (0.001)	0.823 (0.008)
		128	0.983 (0.001)	0.830 (0.011)
		256	0.982 (0.001)	0.821 (0.014)
		512	0.970 (0.008)	0.635 (0.133)
	dbmdz_bert_base_french_europeana_cased	64	0.990 (0.001)	0.898 (0.014)
		128	0.990 (0.001)	0.899 (0.013)
		256	0.990 (0.001)	0.904 (0.006)
		512	0.991 (0.001)	0.912 (0.006)
	dbmdz_bert_base_historic_multilingual_cased	64	0.988 (0.002)	0.884 (0.018)
		128	0.990 (0.001)	0.902 (0.011)
		256	0.990 (0.001)	0.900 (0.008)
		512	0.990 (0.001)	0.900 (0.007)
	xlm_roberta_base	64	0.982 (0.001)	0.806 (0.015)
		128	0.983 (0.001)	0.818 (0.016)
		256	0.983 (0.001)	0.818 (0.020)
		512	0.971 (0.013)	0.593 (0.332)
multi-fr	bert_base_cased	64	0.988 (0.000)	0.877 (0.006)
		128	0.989 (0.001)	0.896 (0.012)
		256	0.989 (0.001)	0.893 (0.007)
		512	0.988 (0.001)	0.885 (0.012)
	bert_base_multilingual_cased	64	0.988 (0.001)	0.878 (0.010)
		128	0.990 (0.001)	0.899 (0.013)
		256	0.990 (0.001)	0.901 (0.005)
		512	0.982 (0.007)	0.817 (0.087)
	dbmdz_bert_base_historic_multilingual_cased	64	0.990 (0.001)	0.898 (0.007)
		128	0.990 (0.001)	0.899 (0.009)
		256	0.990 (0.001)	0.901 (0.013)
		512	0.989 (0.001)	0.893 (0.013)
	xlm_roberta_base	64	0.984 (0.000)	0.831 (0.004)
		128	0.983 (0.001)	0.826 (0.008)
		256	0.982 (0.001)	0.811 (0.009)
		512	0.977 (0.003)	0.744 (0.033)

TABLE E.5

Sentence classification results for the German test set. F-score both for the overall classification and for the class with the "Positives", i.e. those sentences with an agency mention. Experiments were run five times per configuration, the values show the mean and the standard deviation in brackets.

Language (train-eval)	Model	Max. Seq. Length	F1 (Both Classes)	F1 (Agency Class)
de-de	bert_base_cased	64	0.986 (0.001)	0.852 (0.010)
		128	0.985 (0.002)	0.842 (0.020)
		256	0.986 (0.003)	0.852 (0.024)
		512	0.986 (0.001)	0.861 (0.012)
	bert_base_german_cased	64	0.986 (0.002)	0.862 (0.022)
		128	0.989 (0.003)	0.884 (0.028)
		256	0.988 (0.002)	0.872 (0.015)
		512	0.986 (0.003)	0.862 (0.024)
	bert_base_multilingual_cased	64	0.982 (0.007)	0.827 (0.053)
		128	0.986 (0.002)	0.858 (0.021)
		256	0.985 (0.003)	0.848 (0.026)
		512	0.985 (0.001)	0.852 (0.012)
	dbmdz_bert_base_german_europeana_cased	64	0.981 (0.003)	0.814 (0.026)
		128	0.985 (0.002)	0.841 (0.015)
		256	0.986 (0.003)	0.853 (0.024)
		512	0.981 (0.002)	0.812 (0.017)
	dbmdz_bert_base_historic_multilingual_cased	64	0.983 (0.001)	0.829 (0.006)
		128	0.986 (0.001)	0.860 (0.012)
		256	0.985 (0.002)	0.852 (0.016)
		512	0.983 (0.004)	0.830 (0.042)
	xlm_roberta_base	64	0.982 (0.002)	0.803 (0.022)
		128	0.984 (0.002)	0.830 (0.018)
		256	0.984 (0.002)	0.822 (0.029)
		512	0.981 (0.002)	0.773 (0.032)
multi-de	bert_base_cased	64	0.984 (0.001)	0.834 (0.008)
		128	0.985 (0.001)	0.843 (0.012)
		256	0.984 (0.001)	0.840 (0.009)
		512	0.986 (0.002)	0.854 (0.015)
	bert_base_multilingual_cased	64	0.982 (0.002)	0.818 (0.019)
		128	0.985 (0.002)	0.850 (0.014)
		256	0.985 (0.002)	0.844 (0.020)
		512	0.981 (0.012)	0.751 (0.242)
	dbmdz_bert_base_historic_multilingual_cased	64	0.987 (0.002)	0.868 (0.015)
		128	0.987 (0.002)	0.865 (0.017)
		256	0.986 (0.002)	0.860 (0.014)
		512	0.987 (0.004)	0.865 (0.034)
	xlm_roberta_base	64	0.984 (0.002)	0.827 (0.023)
		128	0.986 (0.002)	0.853 (0.021)
		256	0.986 (0.002)	0.843 (0.015)
		512	0.986 (0.002)	0.849 (0.022)

E.4 Error Analysis (Additional Material)

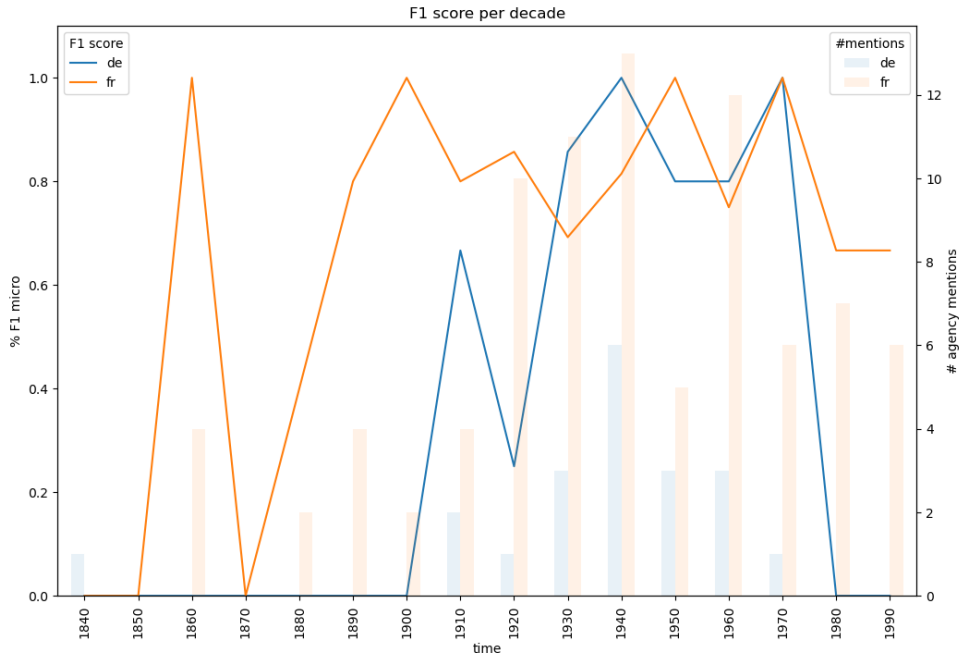


FIGURE E.11

F-score of *bert-base-french-europeana-cased* (fr) and *bert-base-german-cased* (de) for the dev set, split by decade (left y-axis). The bars in the back show the number of mentions existing in the respective test sets per decade (right y-axis).

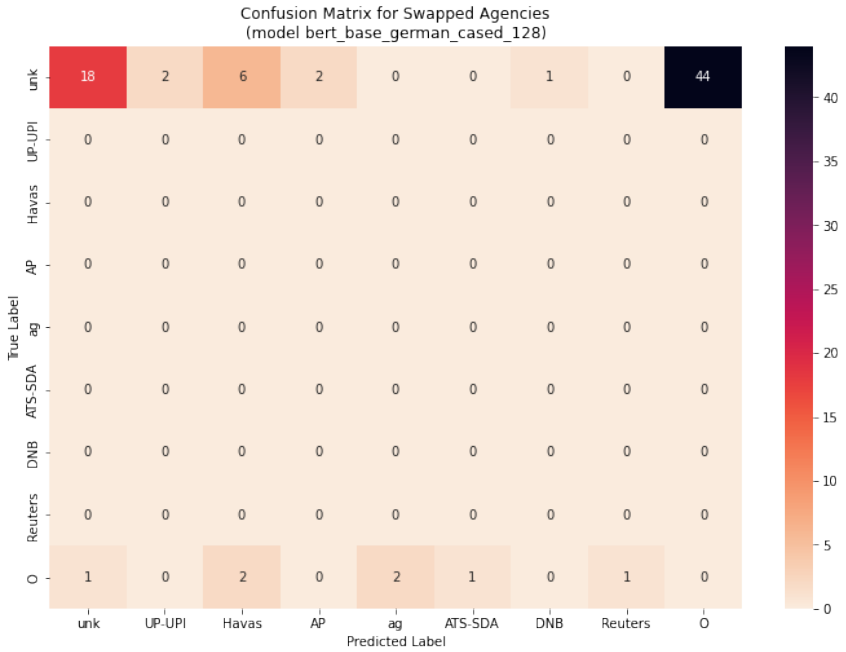


FIGURE E.12

Heatmap showing the distribution over the different classes for the model *bert-base-german-cased* with a maximum sequence length of 128. Results concern the performance on the French test set, where the agencies were swapped with agency names unknown to the model.

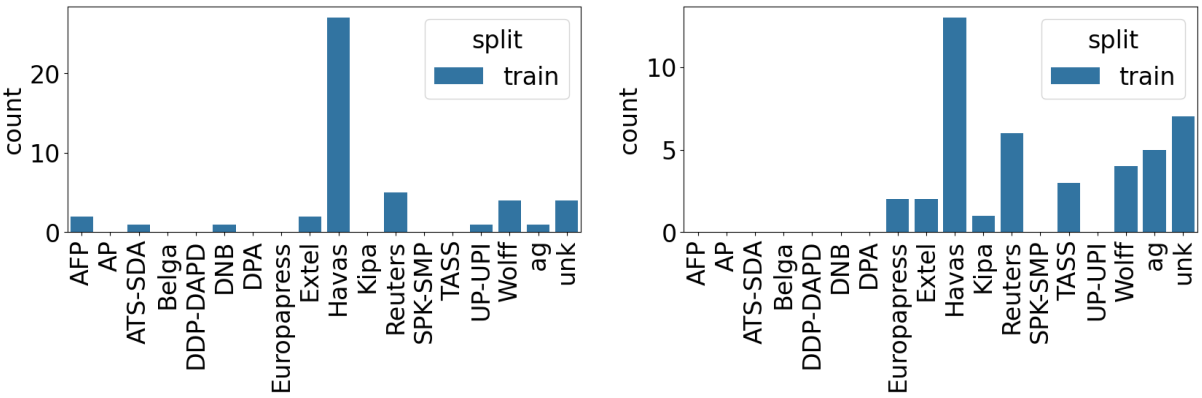


FIGURE E.13

Number of noisy agency mentions in the French (left) and German (right) training sets, split by agency.

F News Agencies in the *impresso* Corpus (Additional Material)

F.1 Quality Assessment

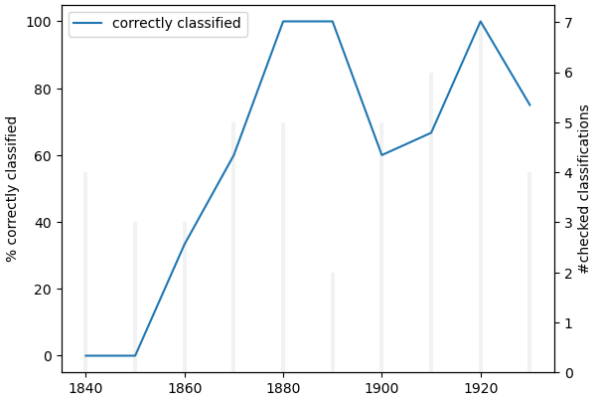


FIGURE F.1
Percentage of correctly classified *Havas* tokens in the checked sample, split by time (left y-axis). The grey bars in the background show the absolute number of checked *Havas* classifications (right y-axis).

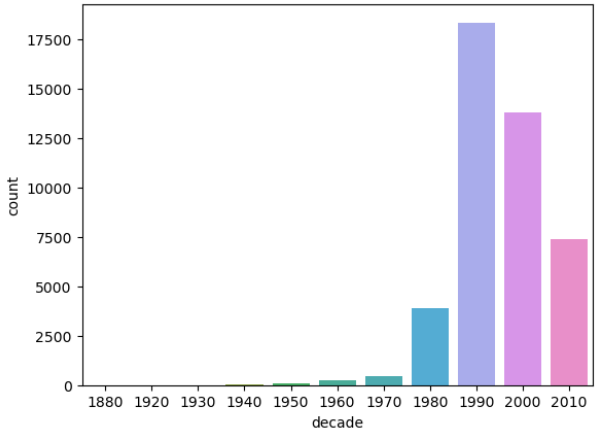


FIGURE F.2
Number of *ATP* tokens classified as agency mentions, split by time. The increase of mentions coincides with the founding of the *Association of Tennis Professionals (ATP)* in 1972.

F.2 News Agencies in the Media Ecosystem

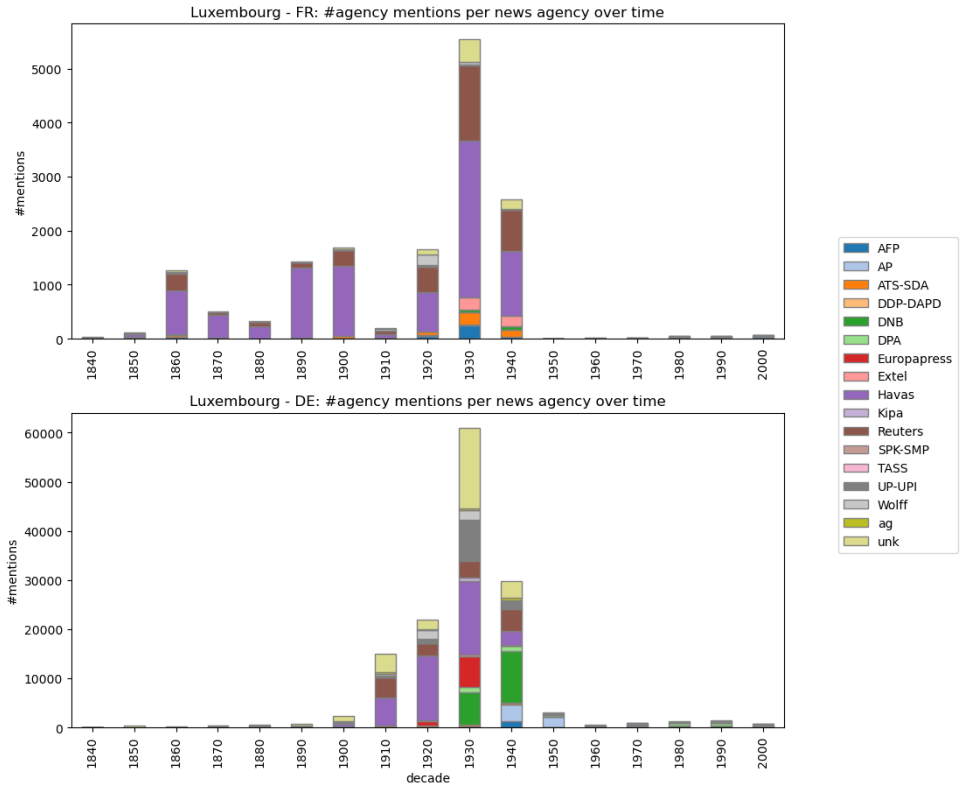


FIGURE F.3
News agency mentions in Luxembourgish newspapers over time, split by language (fr above, de below). Note that the y-axes have different scales.

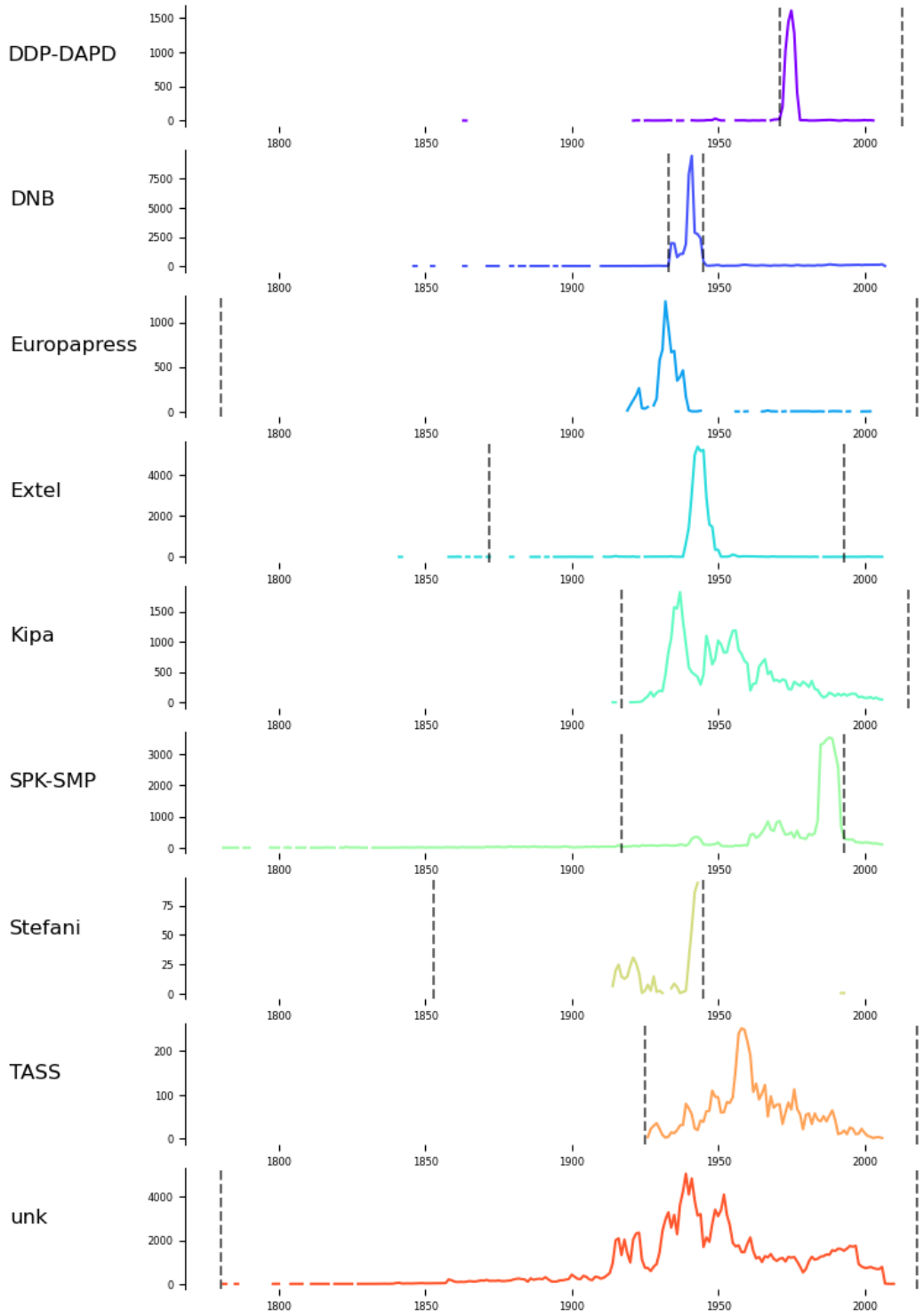


FIGURE F.4

Number of mentions per news agency over time, for a selection of agencies (the other nine agencies can be found in Figure 5.10). The absence of a coloured line means that no agency mentions were detected for the respective year. Each dotted line on the left marks the official founding year of the agency, and on the right dotted line its liquidation. For agencies still present today, the right line was set to 2018, the last year in the *impresso* corpus. Note that the y-axes differ with respect to each agency.