

CoFiF: A Corpus of Financial Reports in French Language

Tobias Daudert^{*†} and Sina Ahmadi[†]

Insight Centre for Data Analytics
National University of Ireland, Galway
{tobias.daudert, sina.ahmadi}@insight-centre.org

Abstract

In an era when machine learning and artificial intelligence have huge momentum, the data demand to train and test models is steadily growing. We introduce CoFiF, the first corpus comprising company reports in the French language. It contains over 188 million tokens in 2655 reports, covering reference documents, annual, semestrial and trimestrial reports. Our main focus is on the 60 largest French companies listed in France’s main stock indices CAC40 and CAC Next 20. The corpus spans over 20 years, ranging from 1995 to 2018. To evaluate this novel collection of organizational writing, we use CoFiF to generate two character-level language models, a forward and a backward one, which we use to demonstrate the corpus potential on business, economics, and management research in the French language.

The corpus is accessible on Github ¹.

1 Introduction

Current research approaches progressively use machine learning and artificial intelligence to derive knowledge from large amounts of data. With natural language processing (NLP) being a crucial part in this progress, knowledge extraction from textual data becomes increasingly important and the underlying fuel, texts, are a sought source. While general corpora exist for many languages such as The British National Corpus [Leech, 1992] for British English, the Corpus of Spoken Professional American English [Barlow, 2000] for American English, or the *Corpus de Français Parlé Parisien des années 2000* (CFPP2000) [Branca-Rosoff *et al.*, 2000] for French, domain-specific corpora are still lacking in many cases. Since transfer learning, particularly language models such as ELMo [Peters *et al.*, 2018] or BERT [Devlin *et al.*, 2018], is currently driving NLP research, large unlabeled corpora play a progressively important role. Examples are the 1 billion word benchmark [Chelba *et al.*, 2013], Wiki-103

[Merity *et al.*, 2016], or CommonCrawl².

Considering the domain of business and economics, especially for English, corpora such as the Wall Street Journal (WSJ) Corpus [Paul and Baker, 1992], the 10-k Corpus [Kogan *et al.*, 2009] and the 8-k Corpus [Lee *et al.*, 2014] are popular examples. However, no corpus to date deals with French texts in the field of economics and finance. Such an absence hinders the progress in applying NLP approaches on the textual data related to the financial sector in francophone countries, particularly France, Canada, Belgium and Switzerland. Hence, we present CoFiF, a corpus aggregating French organizational writing into a source to be analysed in the area of business, economic and management. CoFiF contains documents published by companies which have been part of the *Cotation Assistée en Continu* (CAC) 40³ since 2002. CAC40 contains 40 of the 100 largest companies by market capitalization of the stock exchange in Paris. Furthermore, the CAC40 is France’s main stock index and is dominated by French companies, thus, it can be taken as a representation of French companies in general. In addition, we included companies listed at the CAC Next 20 in the corpus. These companies are the 20 largest ones which are listed following the ones in the CAC40, hence, altogether both indices list the 60 largest French companies. The collected document types provide a comprehensive and factual overview of a company’s shape. In addition, their language can also be consulted in linguistic terms. Previous analyses of company reports for English have shown their effect on the financial markets, for instance, Kogan *et al.* linked 10-k reports to market volatility and Lee *et al.* used 8-k reports to predict stock price movement in terms of [up, down, stay] [Kogan *et al.*, 2009; Lee *et al.*, 2014].

The rest of this paper is organized as follows: we first present previously created French corpora, both general and specific in the field of economics and finance. Following a description of CoFiF in section 3, we evaluate our corpus using a language model in section 4. The paper is concluded in section 5.

^{*}Contact Author

[†] Equal first authors

¹<https://github.com/CoFiF/Corpus>

²<http://commoncrawl.org/>

³<https://www.euronext.com/en/products/indices/FR0003500008-XP/Market-Information>

	CAC40			CAC Next 20			All		
	#Tokens	#Sentences	#Reports	#Tokens	#Sentences	#Reports	#Tokens	#Sentences	#Reports
Annual	20141096	550142	587	2778348	57762	133	22919444	607903	720
Semestral	3988379	78810	410	3302992	63233	254	7291371	142042	664
Trimestral	655991	14091	108	745049	15145	228	1401040	29235	336
Ref. docs.	123238519	3252462	736	33699932	1073180	199	156938451	4325641	935
Total	148023985	3895505	1841	40526321	1209320	814	188550306	5104821	2655

Table 1: Number (#) of tokens, sentences, and reports ordered stock indices and report types.

2 Related Work

There is a plethora of corpora available for the French language, both for general purposes [Abouda and Baude, 2005; Eshkol-Taravella *et al.*, 2010; Content *et al.*, 1990; Guillot *et al.*, 2008; Kunstmann and Stein, 2006] and for specific tasks in NLP. Vincent and Winterstein developed a French corpus for sentiment analysis [Vincent and Winterstein, 2013]. Grabaretal *et al.* targeted reports published in the scientific literature or used in medical education to create a French corpus with clinical cases [Grabar *et al.*, 2018]. Mariani *et al.* presented the NLP4NLP Corpus containing scientific articles published over a period of 50-year in the field of speech and natural language processing in various languages, including French [Mariani *et al.*, 2018]. Mondada *et al.* provided the International Ecological Corpus of French (CIEL) which promotes comparative analysis in the field of linguistic ecology of spoken French in francophone countries [Mondada and Pfänder, 2016]. The Sequoia corpus [Candito and Seddah, 2012] is a syntactically annotated French corpus containing phrases from the French Europarl [Koehn, 2005], l’Est Républicain regional newspaper articles, French Wikipedia, and documents from the European Medicines Agency. Similarly, Martineau *et al.* presented a morphosyntactically structured and annotated corpus (MCVF) to study morphosyntactic variations based on time and social distribution [Martineau, 2008]. Targeting contemporary French, Benzitoun *et al.* [Benzitoun *et al.*, 2016] assembled the ORFÉO which contains 4 million and 6 million words of spoken and written French, respectively.

Regarding corpora in economics and finance for other languages, Kloptchenko *et al.* [Kloptchenko *et al.*, 2004] were the pioneers in producing a corpus based on organizational English content for sentiment analysis for stock market prices prediction. A significant resource is the 10-K Corpus [Kogan *et al.*, 2009] which is composed of 54,379 annual reports in English from 10,492 different companies covering a time interval from 1996 up to 2006. This corpus has paved the way for further tasks in economics and financial text analysis. Similarly, Lee *et al.* created a corpus of 8-k reports which is subsequently used for stock price prediction [Lee *et al.*, 2014]. Recently, Händschke *et al.* [Händschke *et al.*, 2018] introduced the JOCo corpus which contains 5,000 reports (282M tokens) of corporate annual and social responsibility reports from UK, German and US companies for the period of 2000 to 2015.

Despite the need, there have been few efforts in creating French corpora in economics and finance. Verlinde *et al.* targeted Belgian and French newspapers published from 1986

to 1992 to purposely create a corpus containing 1 million words to analyze word combinations used in economical discourse [Verlinde, 1997]. Similarly, Foltête focused on newspaper articles published in France and created a corpus to analyze how economical discourse changes with respect to distributional semantics and transformational grammar [Foltête, 1999]. Focussing on the differences between specialized and non-specialized texts, Cabré presented a multilingual corpus in the field of economics containing 78k words [Cabré, 2007]. Gautier described the construction of a corpus based on the economic articles published in Le Monde and Les Echos for the study of the relation of neologism and economic crises [Gautier, 2012]. Gallego introduces a comparable corpus, COMENEGO, in French and Spanish in business, for translation purposes and a discursive analysis approach based on metadiscourse [Gallego-Hernández, 2013]. According to the literature, no French corpus of significant size is provided in finance and economics so far. To bridge this gap, we introduce the first French corpus dealing with company reports.

3 Corpus Description

Our selection criteria are based on the coherence of the published documents in the field of economics and finance. We can categorize such documents in four types: reference documents (*documents de référence*), which are published annually, usually in the months following the end of the calendar year, and contain information regarding the financial situation and perspectives of a company; annual results (*résultats annuels*) which summarizes a company’s business and activities throughout the previous year; semestrial results (*résultats semestriels*) and trimestrial results (*résultats trimestriels*) which are similar to the annual reports except that they are published every six months and three months, respectively. We included reference documents since some companies also consider this document type as the annual report. Moreover, we found that most companies publish their reference documents on a regular basis. This is not the case for other document types, particularly trimestrial results. All the collected document types report financial results and provide information to financial analysts, institutional investors and individual shareholders. In France, the *Autorité des marchés financiers*⁴ (Financial Markets Regulator), ensures standardization in reporting financial results by requiring companies to follow a certain template.

Regarding the targeted companies, we focus on the CAC40 and the CAC Next 20. Both indices combined contain 60

⁴<https://www.amf-france.org>

of the 100 largest companies by market capitalization of the stock exchange in Paris, thus we found these an appropriate choice for our corpus. Furthermore, we consider reports published between 1995 and 2019.

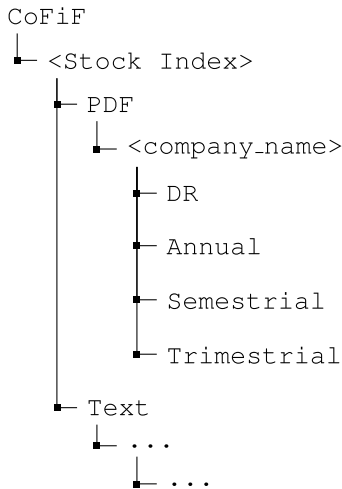


Figure 1: Structure of CoFiF

3.1 Data Retrieval and Corpus Structure

Having a strong and representative shortlist of companies in place, collecting documents was mainly done by consulting company’s website and downloading the reports since 1995. Although there are companies which provide such documents classified by year and type, in some cases the collection could be cumbersome due to lack of organization, website performance or non-continuous publishing (e.g. missing years). In other cases, a company might have changed the name during the last two decades, such as ”Orange S.A.” (previously ”France Télécom S.A.”), or is available under a new company name as outcome of a merge with another company such as the case of ”Engie” (merged ”Gaz de France” and ”Suez”). Under these circumstances, we consulted the Web to find archives of previously published documents, particularly <https://www.bnains.org>, and included documents published under previous company names as well. Although the content of the financial document may be similar in most cases, such similarity is seen to a lesser extent in structure. Companies publish their financial information in various formats ranging from plain text and HTML to tabular and graphical representations. Given the availability of searchable Portable Document Format (PDF), we only included such documents in our corpus. The period of 20 years was chosen based on the availability of reports.

After downloading the reports, we extracted the texts with the command-line software `pdftotext`⁵ in UTF-8. We did not perform further preprocessing on the datasets as we believe that certain tasks, such as document structure extraction, may require different information which should not be affected by the preprocessing step. Nonetheless, we did not observe

⁵<http://www.xpdfreader.com/>

much noise in the collected text. To facilitate document processing, we provide meta-data in the structure of the corpus and the document names. Figure 1 illustrates the structure of CoFiF where documents are classified based on stock indices, CAC20 and CAC40, company name and document types. Further information regarding the publication date is provided in the file name. The structure and the file names in the PDF and Text directories are identical.

3.2 Corpus Analysis

In the following step, we conducted a corpus analysis with the help of the Natural Language Toolkit (NLTK) [Loper and Bird, 2002] for the segmentation and counting of tokens as well as sentences. Table 1 presents the results of the corpus analysis based on the document types and stock indices.

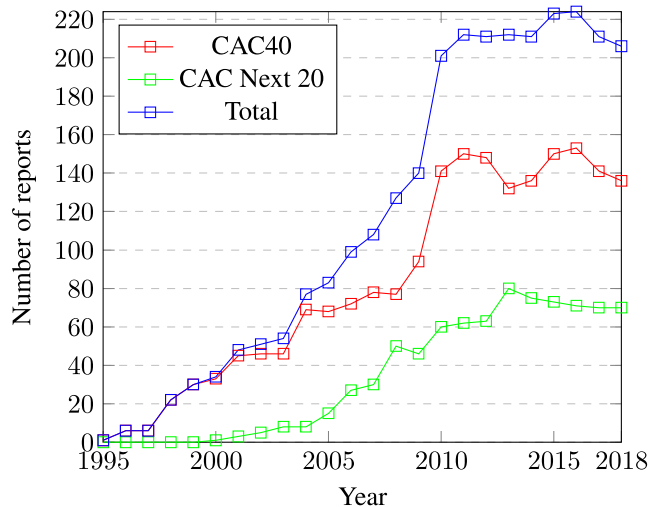


Figure 2: Distribution of reports per year

4 Experiments

To evaluate the corpus and show its potential for NLP tasks in French, particularly within the business and economic domain, we created two character-level language models (LM), trained using a forward recurrent neural network (RNN) and a backward RNN. To prepare the data for the LM, we removed repeated empty lines and breaklines, and aligned the content at the line beginning if necessary. Both language models are generated using a modified version of the NLP library `flair` [Akbik *et al.*, 2018; Akbik *et al.*, 2019]. To train both models we apply the following parameters: `hidden_size` 2048, `n_layers` 1, `sequence_lenght` 250, `mini_batch_size` 100, and `epochs` 3.

In addition to serving as a language model in its natural sense, it also provides word embeddings which can be used in downstream tasks such as text classification or sentiment analysis. The CoFiF word embeddings have shown their use in a sentence boundary detection task which achieved good performance obtaining an F1 score of 0.91 [Daudert and Ahmadi, 2019]. To evaluate both language models, we con-

Sentence	Perplexity
Perspectives d’avenir et principaux risques.	1.7892
Perspectives avenir et principaux risques.	2.9605
Le chiffre d’affaires de l’activité autocars augmente principalement suite à une amélioration du prix moyen, et ce malgré un recul des volumes de 3 %.	1.9745
Le chiffre d’affaires l’activité autocars augmente principalement suite de une amélioration du prix moyen, et ce malgré un recul dès volumes 3 %.	2.9471
Cette stratégie permettrait ainsi d’accroître les péages ferroviaires perçus par Groupe Eurotunnel pour l’utilisation de son infrastructure.	2.4411
Cette stratégie permettrait ainsi d’accroître les péages ferroviaires perçus par Groupe Eurotunnel que l’utilisation son infrastructure.	2.8991

Table 2: Six sample sentences and their perplexity scores retrieved by the character-level forward language model. The upper sentence of each pair is the original sentence, the lower sentence is the modified and wrong sentence.

bénéfice		perte		croissance		impôt		économie	
profit	0.715	dépréciation	0.666	progression	0.857	impôts	0.783	agriculture	0.672
versement	0.544	moins-value	0.599	décroissance	0.782	fonctionnelle	0.606	installation	0.609
solde	0.534	variation	0.571	hausse	0.731	imputation	0.592	énergie	0.603
résultat	0.512	insuffisance	0.515	amélioration	0.719	amortissement	0.535	problématique	0.593
dividende	0.512	diminution	0.505	dynamique	0.716	déduction	0.533	innovation	0.581

Table 3: Five sample word embeddings and their neighbours based on the cosine similarity.

ducted an experiment based on the sentence perplexity. First, we extracted 100 randomly chosen sentences from an annual report external to the corpus. Second, we duplicated these sentences and modified all 100 duplicates forcing grammatical and syntactical incoherences. In the following step, we calculated the sentence perplexity for each of the 200 test sentences. Lastly, we evaluated the correctness of the model’s predictions by using sentence perplexity scores. The model is correct when it returns a lower perplexity score for the initial sentence and a higher score for the modified sentence. The prediction is incorrect When the modified sentence receives the lower score. From the 100 test sentence pairs, our model detected all 100 original sentences correctly. Examples are presented in Table 2.

Furthermore, we trained a Word2Vec model [Mikolov *et al.*, 2013] on the cleaned textual data of CoFiF and evaluated the quality of the retrieved word embeddings. Five sample word embeddings with their neighbours and the respective cosine similarity are shown in Table 3. Looking at the term *économie*, the five neighbours sample suggest that the French *économie* has important *agriculture*, *énergie* and *innovation* sectors. France was the sixth largest agricultural producer in the world and the largest within the European Union in 2011⁶. The research and innovation sector is also important to France with a total spending of 2.26% of the gross domestic product (GDP) leading to the fourth position among all the countries in the Organisation for Economic Co-operation and Development (OECD). The energy (*énergie*) sector plays its role within the French economy (*économie*); France is leading worldwide when it comes to nuclear energy and, as a re-

sult, the smallest emitter of carbon dioxide among the seven largest industry nations⁷. Similarly, one can identify the relatedness of terms such as *bénéfice*(gain), *profit*(profit), *versement*(payment), *solde*(balance), *résultat*(result), and *dividende*(dividend). Hence, we can say that the Word2Vec model can adequately capture the relations between terms in CoFiF.

5 Conclusion

In this paper, we present a novel corpus comprising French annual and semester reports named CoFiF. CoFiF contains a total of 188 million words and is, due to its careful company selection, a good representation for organizational writing in the French business and economic domain. Our preliminary analysis shows CoFiF’s potential to foster business, economic, and management research in the French language. Furthermore, we created two character-level language models which can be used in manifold ways such as the calculation of sentence perplexities or the extraction of word embeddings for downstream tasks. Altogether, this work aims at paving the way for further research in this area which was, until now, hindered by the absence of a publicly available language resource.

Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289, co-funded by the European Regional Development Fund.

⁶<https://web.archive.org/web/20111009235442/http://ambafrance-us.org/spip.php?article511>

⁷https://unstats.un.org/unsd/environment/air_co2_emissions.htm

References

- [Abouda and Baude, 2005] Lotfi Abouda and Olivier Baude. Constituer et exploiter un grand corpus oral: choix et enjeux théoriques. le cas des eslo. In *Corpus en Lettres et Sciences sociales, Des documents numériques à l'interprétation*, 2005.
- [Akbik et al., 2018] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.
- [Akbik et al., 2019] Alan Akbik, Tanja Bergmann, and Roland Vollgraf. Pooled contextualized embeddings for named entity recognition. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page to appear, 2019.
- [Barlow, 2000] Michael Barlow. *Corpus of Spoken, Professional American-English*. Rice University, 2000.
- [Benzitoun et al., 2016] Christophe Benoit, Jeanne-Marie Debaisieux, and Henri-José Deulofeu. Le projet orféo: un corpus d'étude pour le français contemporain. *Corpus*, (15), 2016.
- [Branca-Rosoff et al., 2000] Sonia Branca-Rosoff, Serge Fleury, Florence Lefevre, and Mat Pires. Discours sur la ville. *Corpus de français parlé parisien des années*, 2009, 2000.
- [Cabré, 2007] M Teresa Cabré. Constituer un corpus de textes de spécialité. *Cahiers du CIEL*, pages 37–56, 2007.
- [Candito and Seddah, 2012] Marie Candito and Djamel Seddah. Le corpus sequoia: annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical (the sequoia corpus: Syntactic annotation and use for a parser lexical domain adaptation method)[in french]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 321–334, 2012.
- [Chelba et al., 2013] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Philipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013.
- [Content et al., 1990] Alain Content, Philippe Mousty, and Monique Radeau. Brulex. une base de données lexicales informatisée pour le français écrit et parlé. *L'année Psychologique*, 90(4):551–566, 1990.
- [Daudert and Ahmadi, 2019] Tobias Daudert and Sina Ahmadi. Nuig at the finsbd task: sentence boundary detection for noisy financial pdfs in english and french. In *The First Workshop on Financial Technology and Natural Language Processing (FinNLP 2019)*, Macao, China, 2019.
- [Devlin et al., 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [Eshkol-Taravella et al., 2010] Iris Eshkol-Taravella, Olivier Baude, Denis Maurel, Linda Hriba, Céline Dugua, and Isabelle Tellier. Un grand corpus oral disponible: le corpus d'orléans 1 1968-2012. *Traitement automatique des langues*, 53(2):17–46, 2010.
- [Foltête, 1999] Isabelle Foltête. 11. analyse du discours économique dans le cadre d'une linguistique distributionnelle et transformationnelle. *Modèles linguistiques*, 20(40):119–134, 1999.
- [Gallego-Hernández, 2013] Daniel Gallego-Hernández. Comenego (corpus multilingüe de economía y negocios): a metadiscursive analysis approach. *Procedia-Social and Behavioral Sciences*, 95:146–153, 2013.
- [Gautier, 2012] Laurent Gautier. *Les discours de la bourse et de la finance*, volume 94. Frank & Timme GmbH, 2012.
- [Grabar et al., 2018] Natalia Grabar, Vincent Claveau, and Clément Dalloux. CAS: French corpus with clinical cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 122–128, Brussels, Belgium, October 2018. Association for Computational Linguistics.
- [Guillot et al., 2008] Céline Guillot, Serge Heiden, Alexei Lavrentiev, and Christiane Marchello-Nizia. Constitution et exploitation des corpus d'ancien et de moyen français. *Corpus*, (7), 2008.
- [Händschke et al., 2018] Sebastian GM Händschke, Sven Buechel, Jan Goldenstein, Philipp Poschmann, Tinghui Duan, Peter Walgenbach, and Udo Hahn. A corpus of corporate annual and social responsibility reports: 280 million tokens of balanced organizational writing. In *Proceedings of the First Workshop on Economics and Natural Language Processing*, pages 20–31, 2018.
- [Kloptchenko et al., 2004] Antonina Kloptchenko, Tomas Eklund, Jonas Karlsson, Barbro Back, Hannu Vanharanta, and Ari Visa. Combining data and text mining techniques for analysing financial reports. *Intelligent Systems in Accounting, Finance & Management: International Journal*, 12(1):29–41, 2004.
- [Koehn, 2005] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.
- [Kogan et al., 2009] Shimon Kogan, Dmitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280. Association for Computational Linguistics, 2009.
- [Kunstmann and Stein, 2006] Pierre Kunstmann and Achim Stein. Le nouveau corpus d'amsterdam. *Le nouveau corpus d'Amsterdam. Actes de l'atelier de Lauterbad*, pages 9–27, 2006.
- [Lee et al., 2014] Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, and Dan Jurafsky. On the importance of text analysis for stock price prediction. In *LREC*, pages 1170–1175, 2014.

- [Leech, 1992] Geoffrey Neil Leech. 100 million words of english: the british national corpus (bnc). 1992.
- [Loper and Bird, 2002] Edward Loper and Steven Bird. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*, 2002.
- [Mariani *et al.*, 2018] Joseph Mariani, Gil Francopoulo, Patrick Paroubek, and Frédéric Vernier. The nlp4nlp corpus (ii): 50 years of research in speech and language processing. *Frontiers in Research Metrics and Analytics*, 3:37, 2018.
- [Martineau, 2008] France Martineau. Un corpus pour l’analyse de la variation et du changement linguistique. *Corpus*, (7), 2008.
- [Merity *et al.*, 2016] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.
- [Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [Mondada and Pfänder, 2016] Lorenza Mondada and Stefan Pfänder. Corpus international écologique de la langue française (ciel-f): un corpus pour la recherche comparée sur le français parlé. *Corpus*, (15), 2016.
- [Paul and Baker, 1992] Douglas B Paul and Janet M Baker. The design for the wall street journal-based csr corpus. In *Proceedings of the workshop on Speech and Natural Language*, pages 357–362. Association for Computational Linguistics, 1992.
- [Peters *et al.*, 2018] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.
- [Verlinde, 1997] Serge Verlinde. Le vocabulaire des fluctuations dans le discours économique: synonymie et combinatoire. *Meta: journal des traducteurs/Meta: Translators’ Journal*, 42(1):5–14, 1997.
- [Vincent and Winterstein, 2013] Marc Vincent and Grégoire Winterstein. Building and exploiting a french corpus for sentiment analysis (construction et exploitation d’un corpus français pour l’analyse de sentiment)[in french]. *Proceedings of TALN 2013 (Volume 2: Short Papers)*, 2:764–771, 2013.