Beyond the Salary: Unpacking Job Attributes in French Ads with NLP

Sofia VACA, Alfonso AWADALLA

April 9, 2025

1 Research Question

In this project, we aim to analyse the job attributes advertised by employers in France by exploiting the language used in job vacancies. The main question we aim to answer is "What are the pay-related and non-pay-related job attributes advertised by employers in job ads across France?". Additionally, we examine whether variation in advertised attributes can be explained by the sector and location of the job.

To address these questions, we apply Natural Language Processing (NLP) techniques and a supervised machine learning approach to systematically extract and classify information from the textual content of job ads.

2 Literature Review

A growing strand of the literature explores the value of job attributes beyond wages, showing that workers may value non-wage amenities —such as flexibility, health coverage, or work environment—sometimes even more than a wage increase. Maestas et al. (2023) use stated-preference experiments to estimate workers' willingness to pay for these characteristics, highlighting how non-monetary factors can play a crucial role in job choices and overall job satisfaction.

Other work shifts focus on how employers communicate job attributes. A key reference for our project is Audoly et al. (2024), who use NLP techniques to classify pay and non-pay attributes in Norwegian job ads. They find that while 55% of ads mention wages, nearly all refer to non-wage attributes. Industry and occupation explain only 30–40% of the variation in advertised content, but this rises to 50–70% with firm-level fixed effects—suggesting consistent, firm-specific posting styles. This paper is particularly relevant to our work, as we apply a similar methodology to the French context, where such an analysis has not yet been conducted.

Since there is no unified taxonomy of job attributes across countries, the ILO (2025) proposes a rule-based NLP method to extract non-wage amenities from online ads in emerging economies, adapting U.S.-centric approaches to more diverse labor markets. Their classification includes dimensions like flexibility, career development, and work-life balance—features often overlooked in standard labor statistics. Alongside Audoly et al. (2024), this provides a valuable framework for our own classification strategy.

Other studies have applied NLP to detect specific non-wage amenities. Hansen et al. (2023) focus on flexible work arrangements and identify remote work offers, showing a threefold increase since 2019. Adams-Prassl et al. (2020) focus on employer-provided training, finding higher rates of this amenity in more concentrated labor markets. These studies show how NLP enables scalable analysis of job quality features that are otherwise hard to observe. Similarly, Deming & Kahn (2018) use NLP techniques in job vacancy texts to uncover firm-level variation in skill demands (both cognitive and social). Their findings highlight that job postings can reveal meaningful firm-level variation in skill demands.

Finally, the literature on search frictions and matching theory helps explain why the information in job ads matter. In a labor market with imperfect information, job ad language acts as a key signal. Marinescu &

Wolthoff (2020) find that job titles, more than wages, predict the number and quality of applicants. Sockin & Sojourner (2023) further suggest that jobseekers actively seek employer insights, reinforcing the importance of accessible, firm-level information. These studies support our focus on job ads as a central channel of information in the matching process.

3 Data

Our main data source is the JOCAS (Job Offers Collection and Analysis System), a database created by DARES in response to the rise of online job posting and job search platforms in France. Using a web scraping tool, DARES collects online job advertisements daily from selected French websites. The dataset includes key variables such as detailed job descriptions, job location, ROME codes (the official French job classification system), and other relevant information for each position. We will focus on data from the year 2020, and incorporate additional years if accessible. Our analysis will be limited to metropolitan France.

JOCAS does not cover all occupations equally. Managerial fields or those relying heavily on online recruitment are overrepresented, while sectors with mass recruitment or informal hiring practices are typically underrepresented (DARES, 2023). Additionally, the dataset contains substantial missing data across several variables, which we account for when selecting variables for analysis.

To analyse sectoral differences, we map ROME codes to occupational families using the official correspondence table provided by DARES (2024). Each ROME code is matched to both FAP87 and FAP22 classifications, which group occupations by shared skills or work contexts. FAP22 provides 22 broad families, while FAP87 offers a more detailed breakdown into 87 families, allowing flexible aggregation levels in the analysis.

4 Empirical Strategy

The empirical approach of this paper relies primarily on detailed text analysis of job advertisements and leverages variation in the presence of different types of attributes across job sectors. While skill-related attributes have been extensively studied, resulting in a comprehensive taxonomy (ILO, 2025), amenities have predominantly been explored individually. Audoly et al. (2024) significantly advanced the literature by analysing a broad range of amenities simultaneously.

Building upon the methodology outlined by ILO (2025), we adapt Audoly et al. (2024) taxonomy to our French context. This adaptation involves translating the taxonomy into French, ensuring translation accuracy, and enriching the dictionary with variations specific to French job-market conventions. Additionally, we enhance the comprehensiveness of our attribute dictionary by employing topic modeling (Latent Dirichlet Allocation, LDA) and word embedding techniques on a subset of the JOCAS dataset, ensuring we robustly capture all relevant attributes from our textual data.

4.1 Econometric Framework

To address our research questions, we estimate a series of logistic regression models of the form:

$$\text{AttributePresence}_i = \alpha + \beta \text{Sector}_i + \gamma \text{Location}_i + \epsilon_i, \tag{1}$$

Where $\text{AttributePresence}_i$ is a binary variable indicating the presence of a specific job attribute in ad i, Sector_i indicates the sector classification (FAP87 or FAP22), and Location_i includes controls for geographical variation with potential for additional controls as data availability allows. These models allow us to quantify the extent to which sectoral variation accounts for differences in how job attributes are advertised.

This strategy does not aim to identify causal relationships but rather to descriptively map and explain variation in job attribute advertising using observable sectoral and locational characteristics. Pseudo-R^2 values and regression coefficients will be used to assess explanatory power and effect sizes.

References

DARES. (2024). *La nomenclature des familles professionnelles 2021.*
https://dares.travail-emploi.gouv.fr/donnees/la-nomenclature-des-familles-professionnelles-2021. Accessed March 2025.

ILO. (2025). *Developing a New Method to Uncover Skills Trends in Emerging Economies Using Online Data and NLP Techniques.*
ILO Methodological Brief No. 1. https://doi.org/10.54394/HQQX3200. Accessed April 2025.

Audoly, R., Bhuller, M., & Reiremo, T. (2024). *The Pay and Non-Pay Content of Job Ads.*
First version: July 2024. This version: September 2024.
https://arxiv.org/abs/2407.13204. Accessed April 2025.

ILO. (2025). *Measuring Quality of Employment in Emerging Economies: A Methodology for Assessing Job Amenities Using Online Data.*
ILO Methodological Brief No. 2. https://doi.org/10.2139/ssrn.3957002. Accessed April 2025.

Deming, D., & Kahn, L. (2018). *Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals.*
Journal of Labor Economics, 36(S1), S337–S397. https://doi.org/10.1086/694106. Accessed April 2025.

Marinescu, I., & Wolthoff, R. (2020). *Opening the Black Box of the Matching Function: The Power of Words.*
Journal of Labor Economics, 38(2), 535–568. https://doi.org/10.1086/705880. Accessed April 2025.

Maestas, N., Mullen, K. J., Powell, D., von Wachter, T., & Wenger, J. B. (2023). *The Value of Working Conditions in the United States and the Implications for the Structure of Wages.*
American Economic Review, 113(7), 2007–2047. https://doi.org/10.1257/aer.20210415. Accessed April 2025.

Hansen, S., Lambert, P. J., Bloom, N., Davis, S. J., Sadun, R., & Taska, B. (2023). *Remote Work across Jobs, Companies, and Space.*
NBER Working Paper No. 31007. https://www.nber.org/papers/w31007. Accessed April 2025.

Adams-Prassl, A., Balgova, M., & Qian, M. (2020). *Flexible Work Arrangements in Low Wage Jobs: Evidence from Job Vacancy Data.*
IZA Discussion Paper No. 13691. https://www.iza.org/en/publications/dp/13691. Accessed April 2025.

Sockin, J., & Sojourner, A. (2023). *What's the Inside Scoop? Challenges in the Supply and Demand for Information on Employers.*
Journal of Labor Economics, 41(4), 1041–1079. https://doi.org/10.1086/722936. Accessed April 2025.

DARES. (2023). *Base JOCAS – Document méthodologique résumé.*
https://dares.travail-emploi.gouv.fr/publication/. Accessed April 2025.