Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 83–94
DOI: 10.15439/2024F5992
ISSN 2300-5963 ACSIS, Vol. 41

Exploring Linguistic and Cultural Differences in Online Job Advertisement Analysis for NLP Applications

Lea Grüner
German Federal Institute
for Vocational Education and Training
Email: lea.gruener@bibb.de

Kai Krüger
German Federal Institute
for Vocational Education and Training
Email: kai.krueger@bibb.de
ORCID: 0009-0000-0596-3621

***Abstract*—Online job advertisements (OJAs) have become a significant data source for analyzing labor market dynamics, offering insights into shifts within occupations, industry sectors, skills, and tasks. This paper investigates the cross-lingual and cultural differences in OJAs and their impact on the transferability of Natural Language Processing (NLP) methods and research scope. By analyzing OJAs from Austria, France, Germany, Italy, Spain, the UK, and the US, we point out substantial variations in document length, diversity metrics, syntactic structures, and content features such as salary information. These differences underscore the challenges in applying NLP methods universally across languages and cultures. Our findings emphasize the need for tailored approaches in NLP research and offer a starting point for developing standardized pipelines for analyzing text genres across different languages.**

I. Introduction

ONLINE Job advertisements (OJAs) have garnered significant attention from researchers across various fields, including social sciences, economics, and computational linguistics. Studies utilizing OJAs have explored labor market trends, skill demands, and occupational shifts [1], [2], [3], [4], [5], [6]. Additionally, OJAs offer insights into demographic targeting and potential discrimination, making them critical for research in Human Resources (HR) and gender studies [7], [8], [9]. The digital migration of job ads has enhanced their accessibility, prompting the development of Natural Language Processing (NLP) methods to process and analyze these texts. However, given the diverse linguistic and cultural contexts of OJAs, it is essential to investigate how these differences affect the transferability of NLP methods and research findings.

The main contributions of our exploratory study are:

1) **Cross-Lingual Variation in Job Advertisements**: We identify and quantify[1] substantial differences in language, structure, and content in job advertisements across seven countries, emphasizing the need for localized approaches in NLP research.
2) **Impact on NLP Method Transferability for Downstream Tasks**: The study highlights how cross-lingual and cultural variations in the text genre job ads can hinder the direct transfer of NLP methods, providing tangible evidence and metrics to support this claim. We discuss how this impacts other text genres as well.
3) **Comprehensive Data Collection and Analysis Pipeline**: We present a robust methodology for scraping OJAs from multiple countries using consistent data sources, ensuring high comparability and reliability of the results.

By exploring the form and content variations in OJAs from different countries and languages, this paper aims to identify potential pitfalls in cross-lingual and -cultural research, providing a foundation for more effective and nuanced NLP applications in the context of global applied NLP.

II. Job Advertisements as Research Subjects

In recent years, the interest in job advertisements (job ads) has increased among a diverse group of actors, including researchers from fields such as social sciences and economics, as well as government agencies and private corporations, who use job ads as a data source to gain insights into labor market dynamics. These insights include shifts in occupations [1], [2], industry sectors, skills [3], [4] and tasks [5], [6]. Beyond these aspects, job ads can also provide perspectives on how various demographic groups are targeted or potentially discriminated against, offering critical data for HR or gender studies [7], [8], [9] research. Furthermore, job ads have been used to study how employers attempt to attract candidates, for example by analyzing benefits they offer [10] or the way they present their corporate identity [11], [12]. Research on job ads has been performed in many different countries and on job ads of many different languages, including, among others, all EU states and the UK [13], [14], Brazil [15], Canada [16], China [7], Japan [17], Mexico [15], Taiwan [18] and the United States [19].
While some of the mentioned studies were of qualitative nature, the rise of big data in recent years and the digital migration of job advertisements have significantly enhanced their accessibility as a data source, which motivated the development of NLP methods to structure and analyze these texts. This study aims to investigate how linguistic and cultural

[1]Our code and supplementary plots can be accessed here: https://github.com/TM4VETR/linguistic-cultural-differences-OJAs/

differences in job ads impact the transferability of NLP methods and research findings across different languages and cultures.

III. Job Advertisements as Text Genre

From a linguistic perspective, job ads can be considered a text genre [20], characterized by their communicative purpose [21], the recurrence of situations they address, the discourse community that produces them, and their primary audience [22]. Thus, the actors working with these texts have certain expectations about the texts' content and linguistic characteristics based on their genre knowledge [23]. However, some studies on other text genres have shown that these assumptions might not always apply to all languages and cultures [24], [25], [26], [27], [28], [29], [30], even though in other studies such evidence has not been found [31]. Studying cross-lingual and -cultural differences within a text genre has also gained great attention by researchers from translation studies [20], [32], [33], [34]. In NLP, however, research on cross-lingual text genre has instead focused on getting models to transfer and adapt genre-specific language information from resource rich languages to resource low languages. We argue that practitioners in applied NLP working with specific text genres can benefit from analyzing cross-lingual and cultural differences within their text genre. Our core argument is that researchers may falsely assume that findings derived from one language or cultural context apply universally within the same genre. This can unfold in two ways:

1) **NLP Methods:** Researchers assume their NLP methods for processing job ads are universally applicable. This could include the usefulness/ translation of word lists, structural patterns for rule based systems or the potential, limitations, strengths and weaknesses of Machine Learning with encoder-based models like BERT [35] or auto-regressive models such as GPT [36].
2) **Research Scope:** Given the diverse purposes for which job ads are used, it is unclear whether all information expected to be included in job ads actually is included across countries.

IV. Research question

Derived from the observations described in the previous two chapters, our central research question is: **Can differences in the form or content of OJAs potentially hinder the transferability of NLP methods or research scope?** To answer this question, we perform quantitative analyses on OJAs from different countries and languages. Our analyses focus on different aspects of OJAs and were designed partly to explore the linguistic differences between the data and partly with regard to specific features derived from potential pitfalls. We chose this mixed approach, because we believe both types of analyses might be beneficial to researchers. Analysing cross-lingual or -cultural differences with regard to a specific pitfall can help uncover and overcome it. However, it is impossible for researchers to identify all such pitfalls exhaustively beforehand, which is why using general descriptive analyses of the language data might help uncover additional instances. We describe our exact methods in Section VI and indicate whenever we use a certain analysis with respect to a specific pitfall.

V. Data

Since no publicly available, comparable job ad dataset across multiple languages exists, we decided to collect our own data by scraping[2]. We faced the challenge that country-specific differences could be influenced by factors like occupation or industry sector, acting as confounding variables. Classifying ads into the respective taxonomies would require complex models and normalization, beyond this project's scope. Our exploratory research indicated that different websites target different audiences, affecting the types of employers and employees. To mitigate this, we scraped data from the same website operating across multiple countries. We scraped data from CareerJet[3] from Austria, France, Germany, Italy, Spain, the United Kingdom, and the United States.

We chose these countries specifically for several, mostly practical reasons. Firstly, we aimed to include different countries sharing the same main language to see whether differences were merely linguistic or also cultural. Subsequently, we have chosen two German and two English speaking countries. Then, for our analyses we make use of existing NLP libraries and models. Especially the use of multilingual models when performing tokenization or NLI analyses (see Section VI) required us to limit ourselves to languages that these models have been pre-trained on. Additionally, since we use Zero-Shot models for our analyses, we wanted to focus on languages that our team members had access to, to be able to manually examine the performance of these models for some basic sanity checks. Lastly, the countries mentioned are all rather large and strong in research, which ensured that there would be enough data, and with regard to our research question it is likely that methods for OJA processing will be developed on data from these countries. The choice of countries, however, is somewhat vulnerable as we discuss in Section IX.

Our target was to gather a dataset of 10,000 OJAs from each country. We argue that this number is sufficiently large to provide meaningful insights into the research questions while remaining computationally manageable considering some of the applied analyses are quite resource intensive, for example the vendi-score calculation (Section VI). We scraped slightly more data in case some of the scraping results were corrupted and randomly sampled to get the desired amount. Data scraping for all countries was completed within a week, with one or two days dedicated to each country. Therefore, our data is also comparable with regard to origin time.

In addition to the job ad data scraped from CareerJet, we ran our linguistic analyses on the Wikilingua dataset [37] for comparison where applicable. This dataset contains a compilation

[2]We are working on publishing our data in compliance with data privacy regulations.

[3]https://www.careerjet.com/, accessed April 2024.

of WikiHow[4] articles and summaries in 18 different languages. Using a reference dataset is useful to distinguish between language and genre related features and differences to some extent. We chose the Wikilingua dataset because it is available for all examined languages (English, German, French, Italian, Spanish), easily accessible, and the texts are of suitable length. To run the analyses, 10.000 Wikilingua articles per language are sampled. Noteworthily, the contents of these articles may partly overlap between languages, since the Wikilingua data contains translations of English articles.

VI. Method

We executed an analysis pipeline with each of the seven data splits from Austria, France, Germany, Italy, Spain, the UK, and the USA to compare job ad data across countries and languages. The examined features include structural, linguistic and job-ad related aspects, precisely document length, lexical and structural diversity, Part-Of-Speech (POS) tags, paragraph and list count and length, language detection, and presence of salary information as a content feature. All aspects except the latter three were analyzed using the Wikilingua [37] reference dataset as well. In the following, the specific measures used to determine the aforementioned features are described.

Document length: The mean, median, and standard deviation of document length in tokens were calculated for each country split. To get the length in tokens, we used two different tokenization methods. The first one is simple white-space splitting, with additionally separating symbols and punctuation and counting them as single tokens. Since pre-trained language models such as BERT [35] use Wordpiece tokenization, we used the BertTokenizer for multilingual BERT [35] for splitting documents as well. This results in higher token counts because tokens are split into sub-tokens. Document length in tokens can be highly relevant for processing in pre-trained language models since BERT-like models usually truncate texts after 512 tokens.

Type-token ratio: The type-token ratio (TTR) is a simple measure for lexical diversity of a text (corpus). It is calculated by dividing the number of types (unique tokens) by the number of tokens (all tokens in a text). Since the TTR is sensitive to text length, we used the standardized TTR (STTR) that computes the TTR for each window w of n tokens and averages over all windows W.

$$STTR = \frac{\sum_{w \in W} \frac{count(types)}{count(tokens)}}{count(w)}. \quad (1)$$

The STTR is a value between zero and one. Values close to one indicate high lexical diversity, lower values point to less diversity. In our experiments, the STTR was calculated with a window size of n=1000. For each country corpus, it was calculated using both on white-space tokenized texts, and lemmas obtained from language models from the SpaCy NLP pipeline [38]. Especially for morphologically rich languages like German, lemmas are more insightful for measuring lexical diversity.

Vendi-score: The vendi-score (VS) [39] is a diversity metric for Machine Learning that can be used for a broad range of matrix-based data types. It is defined as

$$VS_k(x_1, ..., x_n) = exp(-\sum_{i=1}^{n} \lambda_i log \lambda_i) \quad (2)$$

where $x_1, ..., x_n$ is a collection of n samples, k is a pairwise similarity function with the kernel matrix K, and $\lambda_i, ..., \lambda_n$ denote the eigenvalues of K/n. Higher values indicate more diversity within a collection of samples.

We calculated the n-gram VS with n ranging from one to three to measure lexical-structural diversity within a data split. Additionally, the embedding VS with contextual embeddings obtained from multilingual BERT [35] was used to assess semantic diversity. The similarity function k is the cosine similarity for both n-gram and embedding VS.

Generally, diversity on different linguistic levels can offer insights into how complex the downstream task is. This could, for example, be used as an indicator of how much evaluation data is required. Examining diversity is particularly important for approaches that are evaluated on the basis of very few manual examples, e.g. prompting [40], other zero- or few-shot methods [41] or synthetic job ads [42].

POS-tags and POS n-grams: To get insights into the linguistic structure of the texts, we used POS-tags obtained using the SpaCy NLP pipeline [38]. For each country, the frequency of POS-tags was determined. If the frequency distribution of tags varies a lot between countries and datasets, this may point to language- or genre-specific differences in job ad data. It is also interesting to examine whether there are significant discrepancies between the two same language country pairs (Germany-Austria and UK-US). In addition, we count n-grams of POS-tags with n=2 and n=3. Again, this could provide evidence for language- or genre-specific contrasts. For instance, the comparison of POS n-grams allows conclusions on syntactic differences, which can be important for analysis systems based on POS-patterns.

Considering the differences in linguistic structures can be relevant specifically for the transferability of syntax based methods. For example, [6] use verb object pairs to extract tasks from OJAs. While this is reasonable for English job ads, an exploratory analysis of German job ads showed that instead of using such verb-object pairs these tasks were often expressed as compounds (e.g. *Kundenberatung*).

Paragraph/ List count and length: Scraping data from CareerJet allowed us to maintain structural HTML information from the website such as linebreaks and listings (ordered and unordered lists). It is worth mentioning, however, that this approach possibly misses out on some listings, since some may not be appropriately formatted in HTML. We created paragraphs based on linebreaks and lists (a list is counted as one paragraph). The paragraphs and lists were counted per document, and the average count per document was calculated for each country split. In addition, we computed the average paragraph length in whitespace-tokens and the average list

[4]https://www.wikihow.com/Main-Page, accessed April 2024.

length in list items.
This analysis might offer insight into structural differences between job ads from different countries. This is relevant, because the structure of job ads is frequently used in the OJA analysis pipeline. For example, researchers performed text zoning to identify different segments within the job ads such that specific aspects like skills were only extracted from relevant segments [43], [44], [45]. By counting the amount of list items we hoped to gain insight into how many skills, tasks and benefits can on average be found in OJAs of a given country. We assumed that lists almost exclusively list one of these three entities, i.e. not mix these entities or contain information about other topics. One downside of our current approach is, however, that it lacks the ability to classify these entities.

Language Detection: We used an XLM-RoBERTa-base pretrained language detection model [46] to identify job ads not written the respective country's main language. The model is trained to distinguish 21 languages, including the five examined (English, French, German, Italian, Spanish). Determining the percentage of foreign language texts and the languages present in a dataset is crucial for applying NLP methods, since many approaches are built for a single particular language. Choosing a multilingual method, such as a suitable pretrained multilingual model, might be necessary in some cases. Social science researchers in particular are usually interested in investigating the entire country, regardless of the language. Moreover, detecting the languages used in job ads could also provide interesting insights into the target group to be addressed by the employer. For instance, if many job ads from a non-english speaking country are written in English, this could point to a particular interest in international candidates.

Salary information: To determine the presence of salary information in a job ad we used two indicators. First, we extracted information from a text field of the website by storing the corresponding HTML-tag information in our database. This information usually consists of precise numbers such as the hourly, monthly or yearly salary. Second, a zero-shot classification approach based on a multilingual mDeBERTa model trained for Natural Language Inference (NLI) [47] is used. Each paragraph of a job ad was used as a premise tested against the hypothesis *'The line contains information on salary.'*. If the model yielded an entailment probability ≥ 0.9, the paragraph was marked and the whole job ad was labeled with *true*. For each country split, the relative number of job ads containing salary information was computed.
We consider our investigation of salary information as a means to test how research scope varies between cultures, as outlined in Section III. Previous studies have leveraged salary data to analyze phenomena such as the impacts of introducing minimum wage [48]. Therefore, researchers in other countries, where the minimum wage has been implemented, might be interested in using this study as a reference.

VII. Results

In this section we present our results. We focus on the findings most relevant to our discussion, supplementary results and plots can be accessed in our repository.

Document length: Figure 1 plots mean and standard deviation of document length based on the two methods described in Section VI. Overall, there are more tokens using the BERT tokenization, which is expected given the subword tokenization. The majority of country-wise proportions are comparable for both tokenization methods, although some discrepancies can be observed. For example, German, Austrian and French texts are proportionally longer with the BERT tokenization.
It is striking that the UK and the US not only have the longest OJAs, but are also the only countries where the job ads are on average longer than the Wikilingua texts. US ads are longest with a mean length of 740 BERT tokens while Italian ads are shortest with a mean length of 369 tokens, followed by Spain. German, Austrian and French OJAs display similar mean lengths around 550 tokens. For all countries except the UK and the US, texts from the Wikilingua reference dataset are notably longer on average. French Wikilingua texts are longest with a mean length of 728 tokens. In general, the Wikilingua data shows less discrepancies between languages than the OJA data does between countries. However, one possible explanation is the content-wise overlap in Wikilingua data.

Diversity Metrics: STTR, n-gram VS, and embedding VS are all metrics indicating different aspects of corpus diversity. Therefore, we use Table I to give an overview of text diversity in our datasets. We can see that consistently OJA data is more diverse in terms of STTR than the respective Wikilingua counterpart. For both VS metrics we observe the opposite. The Wikilingua data is more diverse here. For STTR, the difference is highest for Austria and lowest for Italy. For n-gram VS it is highest for Italy and lowest for Germany, whereas for embedding based VS it is highest for the UK and lowest for Germany.
The within metric differences between countries are generally low to moderate with a few exceptions. Overall, we see that French and Spanish data is less diverse across most metrics compared to other country splits, whereas German, Austrian and Italian OJA data is rather diverse.

TABLE I Comparison of different text diversity metrics across countries, plus diversity metrics of Wikilingua data for reference.

Country	STTR↑		VS_ngram↑		VS_embedding↑	
	OJA	Wiki	OJA	Wiki	OJA	Wiki
AUT	.45	.36	275	331	1.42	1.56
DE	.43	.36	284	331	1.4	1.56
ES	.38	.34	242	379	1.39	1.61
FR	.37	.32	164	311	1.34	1.50
IT	.41	.38	280	542	1.4	1.66
UK	.40	.34	240	397	1.19	1.46
US	.41	.34	234	397	1.21	1.46

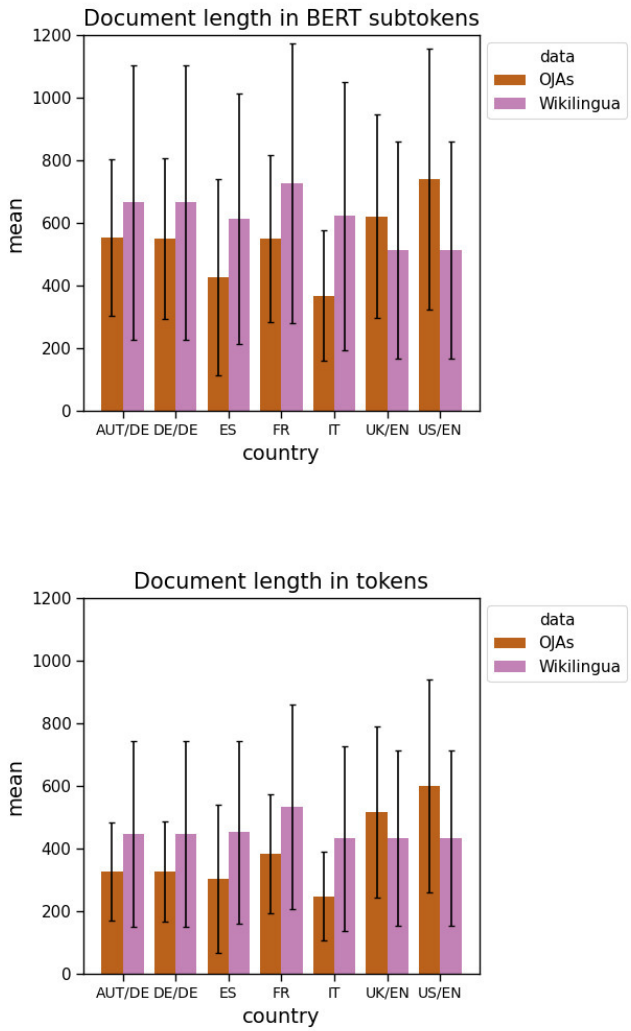


Fig. 1: **Document length.** Upper plot shows mean token count based on BERT subtokens, lower plot based on simple whitespace tokenization. Comparison between countries (x-axis) and datasets indicated by colors.

POS analysis: We analyzed the distribution of different POS types using radar plots as demonstrated in Figure 2 to quickly access the differences between countries with regard to selected POS types. We chose the POS tags ADJ, ADV, AUX, ADP, NOUN, VERB, PRON and PROPN, because we argue that these, mostly containing content words, are most relevant to NLP method design. The most relevant findings from the comparison of OJAs from different countries are:

- In general, for all country splits, there is a major overlap in the distribution of the POS tags examined.
- Nouns have the highest frequencies across all countries.
- Some differences appear to depend on the language family. The Romance languages Spanish, French and Italian have more adpositions (ADP) than the Germanic languages English and German.
- Although the discrepancy is not very profound, data from the English-speaking countries (UK and US) shows an increased occurrence of verbs compared to all other languages. OJAs from German-speaking countries Austria and Germany have the lowest proportion of verbs which may be due to the popularity of substantivations in German.

The corresponding radar plots comparing the two datasets per country can be found in the supplementary material accessible in our repository. The most relevant observations are summarized in the following:

- All countries have more nouns in OJAs than in the Wikilingua data, except from Spain where it is about equal.
- All countries have less verbs in OJAs than in the Wikilingua data.
- All countries have more proper nouns in OJAs than in the Wikilingua data, although the discrepancy is greater for Spain than it is for other countries, like Germany.
- OJAs also tend to have more adjectives and adpositions, whereas Wikilingua data has more auxiliaries and adverbs, although the differences are not very large for the most part. Wikilingua data also has substantially more pronouns.

With regard to bi- or trigrams, the analysis becomes even more complex given the large number of patterns. This makes it difficult to choose individual patterns for comparison. Therefore, we employed Principal Component Analysis (PCA) [49] to identify key syntactic structures that significantly contribute to variations among datasets. PCA effectively reduces dimensionality, transforming the data into principal components that capture the major patterns of variation. This method allows us to highlight the most influential n-grams across different countries. Figures 3a and 3b plot the PCA results across countries. In Table II, we list the five influential bigrams and trigrams identified from the PCA of OJA data. These n-grams have the highest absolute loadings on the first two principal components, indicating their significant contribution to the patterns of variation captured by these components.
Countries sharing the same official language are very close in their principal components. Likewise, there is a clear vertical separation between countries with Romance languages Spain, France and Italy and countries with Germanic languages Austria, Germany, UK and US. This clearly points to similar sentence patterns of the related languages.
When adding the Wikilingua data to the plot, the overlap of each language data was not very pronounced, but rather there was a clear separation between OJA and Wikilingua data. Upon investigating the influential bi- and trigrams per component of the PCA with and without the additional Wikilingua data, we found that the former were much more related to function words rather than content words. This indicates that certain structural patterns in OJA or Wikilingua text genres

govern language specific differences to some extent. On the other hand, the examination of influential bi- and trigrams in OJA data for the components in Figures 3a and 3b reveals that they almost exclusively contain at least one noun. As we have described above, nouns are very frequent in OJAs based on our unigram analysis. However, the exact patterns in which nouns appear seem to vary between languages (see Table II).

TABLE II Influential bigrams and trigrams for the first two principal components.

Component	Influential N-grams
Component 1	Bigrams: ('ADP', 'NOUN'), ('ADJ', 'NOUN'), ('NOUN', 'ADP'), ('NOUN', 'ADJ'), ('DET', 'NOUN') Trigrams: ('NOUN', 'ADP', 'NOUN'), ('ADP', 'NOUN', 'ADP'), ('DET', 'NOUN', 'ADP'), ('ADP', 'DET', 'NOUN'), ('ADP', 'NOUN', 'ADJ')
Component 2	Bigrams: ('DET', 'NOUN'), ('NOUN', 'NOUN'), ('PROPN', 'PROPN'), ('NOUN', 'SPACE'), ('SPACE', 'NOUN') Trigrams: ('NOUN', 'SPACE', 'NOUN'), ('NOUN', 'NOUN', 'PUNCT'), ('NOUN', 'DET', 'NOUN'), ('PROPN', 'PROPN', 'PROPN'), ('ADJ', 'NOUN', 'SPACE')

Paragraph & List information: Table III provides the mean and median amount of paragraphs and lists per country as well as their length. France has the most paragraphs while Austria and Italy have the fewest. US and UK data has the longest paragraphs, and Germany and Italy have the shortest. Likely, these trends are to some extent related to the total lengths of ads per country. However, while having similarly long ads, Germany has substantially more paragraphs than Austria while consequently the paragraphs in Austrian ads are longer.

With regard to lists the most prominent observation is that in Spain, France and Italy more than half of the ads did not have a single list (median = 0). On the other hand, Austria, Germany, the UK and the US had around 2 lists per document on average. However, the average amount of items per list was rather similar across all countries with around 5 items per list as median and average.

Language Detection Overall, the amount of job ads labeled with a language other than the countries' main language was small (Figure 4). Spain had a substantially larger amount compared to the other countries, albeit still minor with around 2% of the texts being labeled with a language other than Spanish. The foreign language texts detected for Spain were exclusively labeled with English and Portuguese. In the other non English-speaking countries, the majority of the foreign texts were detected as English. In the US, the predominant foreign language was Spanish, and Arabic in the UK. Noteworthily, the language detection also revealed a few job ads only contained a very short text, such as a city name and a postal code. The language detection model did not properly work for such instances, resulting in misclassification of the respective samples. Arguably, these texts cannot actually be counted as real job ads. However, since only a very small number of samples was affected, we decided to not further address this issue.

Salary Information: Figure 5 shows the comparison of salary information obtained from a structured website field vs. from the texts. It is apparent that all countries have salary information in the texts that is not included in the structured information of the website. However, the ratio differs significantly. While Germany displays a discrepancy of almost 40 percentage points, in the UK only about 10% of the ads contain additional salary information in text. Generally, Italy has the lowest share of salary information in job ads, whereas UK, US and Austria have a relatively large proportion across both identification strategies.

VIII. Discussion

In this section, we closely examine the results from Section VII, interpreting the outcomes of each method individually (Section VIII-A). We provide explanations for observed phenomena, sometimes overlapping with Section IX, where we discuss limitations of specific methods. However, Section IX focuses more on macro-level limitations rather than individual methods. In Section VIII-B, we adopt a broader perspective, relating our findings to our research question and reflecting on the usefulness of our experiments in gaining insights.

A. Interpreting the Results

Document length: One important finding of looking at the document length is that indeed OJAs frequently exceed the token limit of 512 tokens typically used by transformer-based language models, although for some countries like Italy and Spain, the job ads are exceptionally short and on average below this limit. The large variations of document lengths observed in OJAs (ads in the US are more than twice as long on average) is much more profound than for the Wikilingua data. This indicates that job ads as a text genre can have culture-specific variations that cannot solely be attributed to characteristics of the language.

Several explanation factors are plausible. For example, some cultures might prefer brevity and directness in communication whereas in others there may be a preference for more detailed and comprehensive communication. Another factor could be different price structures for job ad platforms. We know that some job ad platforms have pricing based on text length, and this can differ dependent on the country where the job is

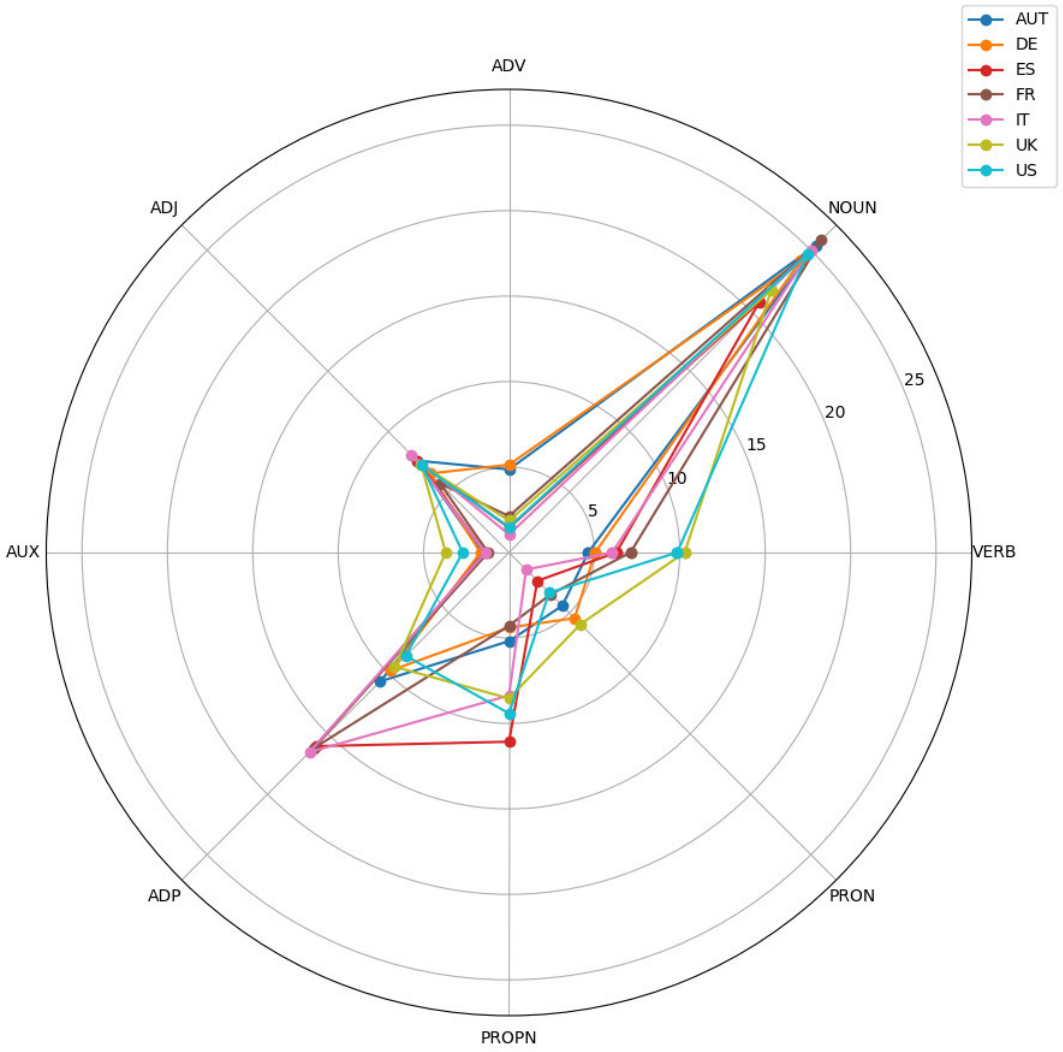
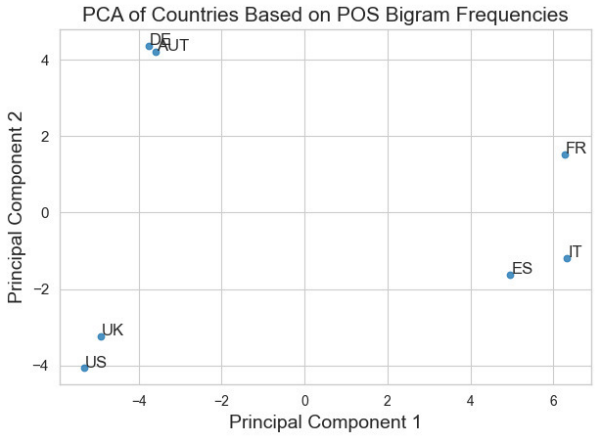


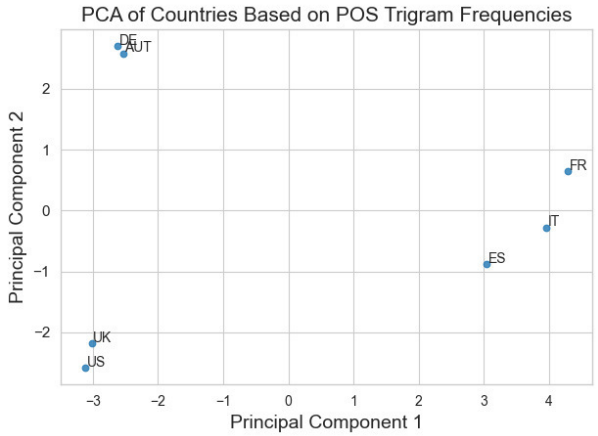
Fig. 2: **POS frequency radar plot.** Radar plot of selected POS types. The plot shows the share of the POS types in the seven OJA country splits.Values in the centre indicate lower frequency of the corresponding POS tag in a country split, values at the edge indicate higher frequency.

TABLE III Comparison of rounded mean and median for the amount of paragraphs per document, length of each paragraph, amount of lists per document, and amount of items per list across countries.

Country	Mean Par./Doc.	Median Par./Doc.	Mean Tokens/Par.	Median Tokens/Par.	Mean Lst./Doc.	Median Lst./Doc.	Mean Items/Lst.	Median Items/Lst.
AUT	13.6	12	23.5	7	2.3	3	5.3	5
DE	20.1	17	16	6	1.8	2	5.2	5
ES	16	13	19	12	0.9	0	5.1	5
FR	21.2	20	18.1	12	0.7	0	5.5	5
IT	14.3	13	17.2	9	0.7	0	4.2	4
UK	18.8	17	27.2	14	2.0	2	6	5
US	20.7	17	28.6	12	2.4	2	6.2	5



(a) Principal Component Analysis of Bigrams (OJA data).



(b) Principal Component Analysis of Trigrams (OJA data).

Fig. 3: **Principal Component Analysis (PCA).** Results for bigrams and trigrams in OJA data. The first two components derived from bigram and trigram frequencies across countries are shown.

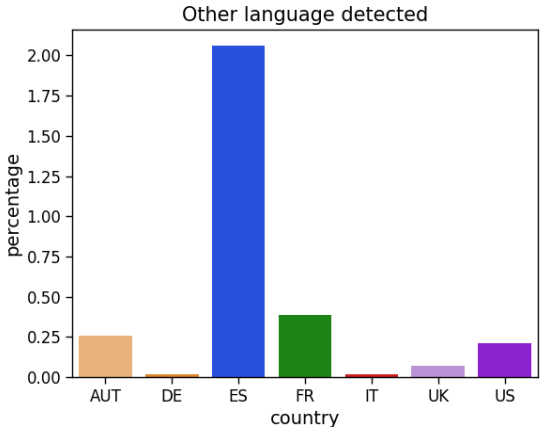


Fig. 4: **Other language detected.** Comparison of the relative share of ads where a language that is not the countries main language was detected.

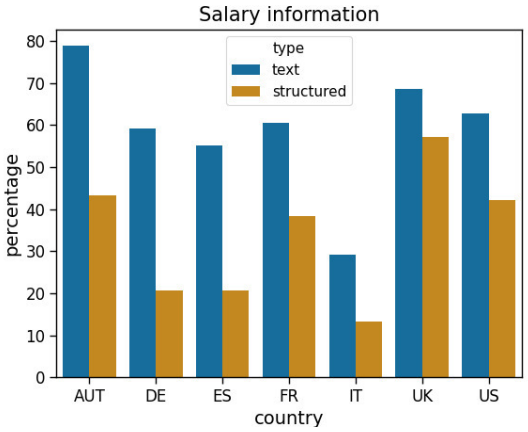


Fig. 5: **Salary information.** Comparison of the relative share of ads with salary information across countries and the two methods of detecting salary.

offered. We do not know as to whether this is the case for our CareerJet data. Even if not, it still might have indirect impact, given that employers may write their ads for several platforms. Another factor could be how formalized qualifications are in a country. For example, Germany has a very formalized vocational education system in place. Therefore, if a company mentions that a potential employee is expected to bring a finished apprenticeship training as, for example, an electrician with a certain specialization, a lot of their expected skills are presumed. In a country with a less formalized vocational education system, the company might feel the need to elaborate more on what exactly they are looking for. Other factors might include legal aspects, labor market dynamics (e.g. what type of occupations are in demand?) or awareness for SEO-optimization.

Diversity Metrics: Our results in Table I show that text diversity differs substantially across countries, but job ads are generally more diverse on a purely lexical level than on other (semantic, syntactic) levels. This indicates that job ads as a text genre have specific linguistic properties. At the same time, the differences between countries, for example between the n-gram VS of France and Germany are quite strong. Also, no country is consistently highest or lowest in all diversity rankings, showing that OJAs from different countries have different properties with regards to various linguistic levels. This leads us to conclude that researchers should carefully reflect the lexical, semantic and syntactical structures of their OJA data when designing research projects. More diverse data might require additional evaluation to ensure robust conclusions.

Syntactic analysis: The analysis of selected POS as well as the PCA analysis of bi- and trigrams show that OJAs can

be seen as a text type with unique syntactic characteristics across countries, such as the increased amount of nouns we found in most countries compared to the Wikilingua data. However, our analysis also reveal that these characteristics do not hold for all countries, e.g. US job ads did not have substantially more nouns than English Wikilingua data. This points to the need of reflecting NLP methods when trying to transfer methods or models from one language to another. Certain patterns, that are typical for entities in one country, might be expressed differently in another country. At the same time, researchers cannot rely solely on knowledge they have about the characteristics of the different languages used in these countries, because differences may be specific to the text genre OJA.

Paragraph & List information: The variability in job ad structure indicated by paragraph amount and length impacts NLP tasks such as text zoning for information extraction. Training NLP models predominantly on data from one country may reduce their effectiveness when applied to structurally different ads from other countries. Moreover, the notable variation in list usage across countries like Spain, France, and Italy, where lists often are absent, complicates the application of list-based analytical methods developed in countries like Germany or the UK. The high similarity in the amount of list items indicates that the amount of skills, tasks or benefits listed in OJAs does not differ substantially across countries.

However, our detection of lists and paragraphs is exclusively based on HTML-analysis. It is possible that a text may simply contain line-breaks combined with list-indicating symbols (like hyphen) as a list. We do, however, know that the website we scraped the data from offers a bullet-point-button in the field where employers put the main body of the job ad.

Language Detection: Our results indicate that ads in other languages than the countries' official language do not occur equally frequent across languages. Based on our results, in Spain this would be an influential factor in the OJA pipeline, possibly requiring an additional data cleaning step or a multilingual approach. At the same time, when having a much larger dataset, this would also provide additional research opportunities, because the occurrence of ads in different languages could be related to other factors. For example, OJAs with different languages might differ regionally or with regard to job requirements.

However, we found several uncertainties with our method. First, the model was limited to the languages it knows. We found that some texts from the Spanish dataset labeled with Spanish or Portuguese are actually written in Catalan, which the model was not trained on. Also, some ads appear to contain very short and noisy texts. If a data point consists of mostly noise, such as symbols or contact addresses, etc., the model tended to predict a completely unrelated language. These cases could easily be identified as false classifications in a visual inspection. However, these instances are misleading in quantitative analysis. Finally, ads might include two (or more) languages [22] to cater to a local as well as international audience. We found that sometimes the ad was repeated in another language. A more refined approach could detect such instances.

Salary Information: With regard to salary information there were major differences between the countries for both structured and text-based information. This shows that transferring research on the basis of correlating salary information to other characteristics from one country to another is not straightforward and requires careful consideration. Also, our method only gives sparse information about the type of salary information detected in texts. Firstly, we do not know if the structured information (e.g. hourly wage) is always repeated in the free text. It might be plausible for the employer to omit repeating it, because they know that employees get that information trough the text field and the visual aid of the websites' structure. We do, however, know that information is repeated at least in some cases, because for some countries the sum of text and structured information is above 100%.

Furthermore, we suspect that there are two major types of information that our NLI model detects except from precise salaries based on exploratory qualitative analysis. The first group is about collective wage information. Our expertise is mostly in the German labor market and here our intuition is that employers rarely mention concrete numbers like an hourly wage, but will mention, if the position is paid based on a collective wage. Especially jobs in the large public sector always include this type of information. The second group is employers advertising their payment using phrases like "attractive" or "above-average" salary. Differentiating between these (and possible more) groups of payment information might give further insight into how the differences between countries can be explained. For example, we suspect the large increase of textual compared to structured salary information in German job ads to be mostly caused by the great amount of collective wage mentions in Germany. Knowing what type of salary information is present in a country and to what extent might further help researchers develop or discard research ideas using salary information obtained from job ads.

B. Reflection on Experiments

Our central research question was whether or not we could find differences in form or content in OJAs that could hinder the transfer of NLP methods or research scope in a major way. Based on our exploratory analysis, we can confirm that there are substantial differences between OJAs from different countries. This applies to both linguistic and content features. Also, the linguistic differences we detected did only partially correspond to the respective cross-lingual differences observed in the Wikilingua reference data. This indicates that there are differences between countries that are specific to the text genre OJA. Interestingly, OJA data from countries that share the same official language had rather similar linguistic features. For example, English texts are the shortest in Wikilingua. However, US OJA texts are longest of all countries, indicating that long texts are not a property to English, but only to US OJAs, indicating cultural factors. Yet, we observe the same (although not quite as extreme) for data from the UK. This

raises the question whether this behavior of the language pairs Germany/Austria and UK/US can be explained by cultural similarities of these countries that share a language or other factors we did not uncover.

With regard to the different methods used, we find that experiments motivated by specific questions, for example text length ("Can we use a 512 token truncation?") or salary information ("Can we correlate salary information to other properties of job ads?") made it easier to draw more concrete conclusions and derive actionable recommendations. For less concretely motivated methods such as text diversity, where we argued that it might be an indicator for how complex various NLP tasks might be (for example, "How many examples do we need to manually evaluate in Zero Shot scenario?"), we cannot simply derive answers based on our results. At the same time, the differences for some of these metrics proved to be quite large. This has implications on two major levels.

Firstly, NLP practitioners developing OJA analysis pipelines should be aware that OJAs can have substantial differences that may go beyond the linguistic differences expected for different languages. If applicable, our results can be used directly to draw conclusions on methods to be applied. Otherwise, our analysis pipeline can be used to analyze other data and then relate to our findings to check whether there may be problems in applying existing NLP OJA methods. Ideally, researchers should add further experiments for the concrete problem they are facing. This prerequisites that they find ways to quantify text properties in a manner that is useful to their problem, which may be challenging.

Secondly, with the rise of NLP methods in various contexts like legal texts, medical texts, social media, literature and so on, we will find an increasing amount of text genres that NLP methods are being developed for. Consequently, NLP practitioners from text genres will look into building upon published work from other languages or countries. As we have shown, some pitfalls exist in this process. Therefore, we advocate researchers working with other text genres to perform analyses similar to ours, using easily accessible tools to gain quantitative insight into their data and how it behaves compared to texts from the same genre in other languages or countries. Ideally, future work would reflect further upon the specific experimental methods and refine a more standardized, yet flexible pipeline that researchers can revert to when they intend to perform an analysis like ours. In this sense, our paper provides a starting point for researchers looking to develop a similar pipeline.

IX. Limitations

Despite our promising results, there are limitations to our study, which we want to discuss in this section. As explained in Section V we chose to scrape data from one website that operates in multiple countries to minimize biases introduced by confounding variables based on target audiences from different job portals. At the same time this choice also means that we equated the properties of the text genre OJA from a given country with OJA data from only a single source. So, while our approach helps in comparability between countries, the countries themselves are not thoroughly represented. Ideally, future work reproduces our experiments with a dataset from mixed source websites that at the same time ensures comparability by choosing similar distributions of relevant variables like occupation or industry sector. Perhaps, this would require an even larger dataset, which, however, makes the experiments more computationally expensive.

Furthermore, there are two factors that are partly of ethical nature. In our study we mentioned that we chose two country pairs that share the same official language in order to better differentiate whether results were based simply on language factors or on cultural factors. That way, we equated a different country to a different culture. This is clearly a very simplified view of the intricacies of cultures and states. We are aware, for example, that within one country, major cultural differences may exist. Also, linguistic discrepancies between a country's regions can be strong, one example being the Catalan ads we found in our Spanish dataset. This may also lead to limited replicability of our experiments for countries where multilingualism is even wider spread. Somewhat related to this issue is the choice of countries, which was mostly based on practicability. We are aware of issues in the NLP community with regards to underrepresented languages [50]. The languages we investigated here all belong to the better represented ones. This is particular important due to the fact that especially underrepresented languages logically have a greater need to adopt methods developed originally for other languages and would therefore profit most from our research. However, due to the reasons explained in Section V including these languages was beyond the scope of this paper. Future research should therefore focus on including underrepresented languages to our analysis.

X. Conclusion

Our analysis of OJAs across various countries and languages reveals substantial linguistic and content-related differences, emphasizing the complexity of transferring NLP methods and research findings across different contexts. The variations in document length, diversity metrics, syntactic structures, and salary information highlight the need for tailored approaches in NLP research. While our study offers valuable insights, it also points to the necessity of further research, particularly involving underrepresented languages and larger, more representative datasets. By acknowledging these differences and adapting NLP methods accordingly, researchers can improve the accuracy and relevance of their analyses in the context of global labor markets. Our findings serve as a foundation for developing standardized yet flexible pipelines for analyzing text genres across different languages and cultures.

References

[1] A. Lima, B. Bakhshi *et al.*, "Classifying occupations using web-based job advertisements: an application to stem and creative occupations," *Economic Statistics Centre of Excellence Discussion Paper*, vol. 8, 2018.

[2] A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica, and A. Seveso, "Neo: A system for identifying new emerging occupation from job ads," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 18, 2021, pp. 16 035–16 037.

[3] E. Senger, M. Zhang, R. van der Goot, and B. Plank, "Deep learning-based computational job market analysis: A survey on skill extraction and classification from job postings," 2024.

[4] M. Buchmann, H. Buchs, F. Busch, S. Clematide, A.-S. Gnehm, and J. Müller, "Swiss job market monitor: A rich source of demand-side micro data of the labour market," *European Sociological Review*, vol. 38, no. 6, pp. 1001–1014, 2022.

[5] E. Atalay, P. Phongthiengtham, S. Sotelo, and D. Tannenbaum, "The evolution of work in the united states," *American Economic Journal: Applied Economics*, vol. 12, no. 2, pp. 1–34, 2020.

[6] E. Atalay, S. Sotelo, and D. Tannenbaum, "The geography of job tasks," *Journal of Labor Economics*, 2023.

[7] P. Kuhn and K. Shen, "Gender discrimination in job ads: Evidence from china," *The Quarterly Journal of Economics*, vol. 128, no. 1, pp. 287–336, 2013.

[8] D. Gaucher, J. Friesen, and A. C. Kay, "Evidence that gendered wording in job advertisements exists and sustains gender inequality." *Journal of personality and social psychology*, vol. 101, no. 1, p. 109, 2011.

[9] S. Chaturvedi, K. Mahajan, and Z. Siddique, "Words matter: Gender, jobs and applicant behavior," *Jobs and Applicant Behavior (February 18, 2024)*, 2024.

[10] M. Ganesan, S. P. Antony, and E. P. George, "Dimensions of job advertisement as signals for achieving job seeker's application intention," *Journal of Management Development*, vol. 37, no. 5, pp. 425–438, 2018.

[11] B. Bullinger, "Companies on the runway: Fashion companies' multimodal presentation of their organizational identity in job advertisements," in *Multimodality, meaning, and institutions*. Emerald Publishing Limited, 2017, vol. 54, pp. 145–177.

[12] J. Binnewitt and T. Schnepf, "Join us to turn the wor (l) d greener!—investigating online apprenticeship advertisements' reference to environmental sustainability," *Zum Konzept der Nachhaltigkeit in Arbeit, Beruf und Bildung—Stand in Forschung und Praxis*, 2022.

[13] Cedefop, *Online job vacancies and skills analysis – A Cedefop pan-European approach*. Publications Office, 2019.

[14] P. Descy, V. Kvetan, A. Wirthmann, and F. Reis, "Towards a shared infrastructure for online job advertisement data," *Statistical Journal of the IAOS*, vol. 35, no. 4, pp. 669–675, 2019.

[15] A. T. S. Calazans, R. A. Paldes, E. T. S. Masson, I. S. Brito, K. F. Rezende, E. Braosi, and N. Pereira, "Software requirements analyst profile: A descriptive study of brazil and mexico," in *2017 IEEE 25th International Requirements Engineering Conference (RE)*. IEEE, 2017, pp. 204–212.

[16] L. Bowker, "What does it take to work in the translation profession in canada in the 21st century? exploring a database of job advertisements," *Meta*, vol. 49, no. 4, pp. 960–972, 2004.

[17] D. Rear, "Converging work skills? job advertisements and generic skills in japanese and anglo–saxon contexts," *Asian Business & Management*, vol. 12, pp. 173–196, 2013.

[18] C.-H. Chung and L.-J. Chen, "Text mining for human resources competencies: Taiwan example," *European Journal of Training and Development*, vol. 45, no. 6/7, pp. 588–602, 2021.

[19] M. A. Kennan, F. Cole, P. Willard, C. Wilson, and L. Marion, "Changing workplace demands: What job ads tell us," in *Aslib Proceedings*, vol. 58, no. 3. Emerald Group Publishing Limited, 2006, pp. 179–196.

[20] A. Trosborg, "Text typology: Register, genre and text type," *Benjamins Translation Library*, vol. 26, pp. 3–24, 1997.

[21] C. L. Engstrom, J. T. Petre, and E. A. Petre, "Rhetorical analysis of fast-growth businesses' job advertisements: Implications for job search," *Business and professional communication quarterly*, vol. 80, no. 3, pp. 336–364, 2017.

[22] F. van Meurs, B. Planken, H. Korzilius, and M. Gerritsen, "Reasons for using english or the local language in the genre of job advertisements: Insights from interviews with dutch job ad designers," *IEEE Transactions on Professional Communication*, vol. 58, no. 1, pp. 86–105, 2015.

[23] C. Berkenkotter and T. N. Huckin, *Genre knowledge in disciplinary communication: Cognition/culture/power*. Routledge, 2016.

[24] Y. Sun, "Genre formation in contexts: a cross-lingual comparison of english ma thesis introductions." *Linguistics & the Human Sciences*, vol. 10, no. 3, 2014.

[25] B. Melander, "Culture or genre? issues in the interpretation of cross-cultural differences in scientific papers," *Genre studies in English for academic purposes*, vol. 9, pp. 211–226, 1998.

[26] Y. Zhu, "A situated genre approach for business communication education in cross-cultural contexts," in *The Routledge handbook of language and professional communication*. Routledge, 2014, pp. 26–39.

[27] C. C. Nickerson, "The usefulness of genre theory in the investigation of organizational communication across cultures," *Document Design*, vol. 1, no. 3, pp. 203–215, 1999.

[28] D. Kuhi and M. Mojood, "Metadiscourse in newspaper genre: A cross-linguistic study of english and persian editorials," *Procedia-Social and Behavioral Sciences*, vol. 98, pp. 1046–1055, 2014.

[29] H. Marefat and S. Mohammadzadeh, "Genre analysis of literature research article abstracts: A cross-linguistic, cross-cultural study," *Applied research on English language*, vol. 2, no. 2, pp. 37–50, 2013.

[30] L. Filipović, "The role of language in legal contexts: A forensic cross-linguistic viewpoint," *Law and Language: Current Legal Issues*, vol. 15, no. 19, pp. 328–343, 2013.

[31] H. Ansary and E. Babaii, "A cross-cultural analysis of english newspaper editorials: A systemic-functional view of text for contrastive rhetoric research," *RELC Journal*, vol. 40, no. 2, pp. 211–249, 2009.

[32] I. G. Izquierdo and V. M. i Resurrecció, "Translating into textual genres," *Linguistica Antverpiensia, new series–themes in translation studies*, vol. 1, 2002.

[33] V. Montalt, P. Ezpeleta-Piorno, and I. García-Izquierdo, "The acquisition of translation competence through textual genre," 2008.

[34] Ł. Biel, "Genre analysis and translation," in *The Routledge handbook of translation studies and linguistics*. Routledge, 2017, pp. 151–164.

[35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[36] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[37] F. Ladhak, E. Durmus, C. Cardie, and K. McKeown, "WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4034–4048.

[38] SpaCy, https://spacy.io/models, Accessed April 2024.

[39] D. Friedman and A. B. Dieng, "The vendi score: A diversity evaluation metric for machine learning," *Transactions on Machine Learning Research*, 2023.

[40] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[41] I. Beltagy, A. Cohan, R. Logan IV, S. Min, and S. Singh, "Zero- and few-shot NLP with pretrained language models," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, L. Benotti, N. Okazaki, Y. Scherrer, and M. Zampieri, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 32–37.

[42] A. Magron, A. Dai, M. Zhang, S. Montariol, and A. Bosselut, "Jobskape: A framework for generating synthetic job postings to enhance skill matching," *arXiv preprint arXiv:2402.03242*, 2024.

[43] A.-S. Gnehm, "Text zoning for job advertisements with bidirectional lstms," 2018.

[44] A.-S. Gnehm and S. Clematide, "Text zoning and classification for job advertisements in german, french and english," in *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, 2020, pp. 83–93.

[45] J. Hermes and M. Schandock, "Stellenanzeigenanalyse in der qualifikationsentwicklungsforschung," *Die Nutzung maschineller Lernverfahren zur Klassifikation von Textabschnitten. Bundesinstitut für Berufsbildung, Bonn*, 2016.

[46] L. Papariello, "xlm-roberta-base-language-detection (revision 9865598)," 2024.

[47] M. Laurer, W. Van Atteveldt, A. Casas, and K. Welbers, "Less annotating, more classifying: Addressing the data scarcity issue of supervised

machine learning with deep transfer learning and bert-nli," *Political Analysis*, vol. 32, no. 1, pp. 84–100, 2024.
[48] E. Andrieu and M. Kuczera, "Minimum wage and skills: Evidence from job vacancy data," The Productivity Institute, Tech. Rep., 2023.
[49] I. T. Jolliffe, *Principal component analysis for special types of data*. Springer, 2002.
[50] J. Nee, G. M. Smith, A. Sheares, and I. Rustagi, "Linguistic justice as a framework for designing, developing, and managing natural language processing tools," *Big Data & Society*, vol. 9, no. 1, p. 20539517221090930, 2022.