ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF TRANSLATION AND INTERPRETING

SECOND CYCLE DEGREE IN TRANSLATION AND TECHNOLOGY (LM-94 CLASS)

WORK AND EMPLOYMENT BEYOND CONVENTIONAL TRANSLATION AND LINGUISTICS: A FOLLOW-UP TO THE UPSKILLS NEEDS ANALYSIS

Dissertation in corpus linguistics

Supervisor

Prof. Silvia Bernardini

Co-Supervisor

Prof. Adriano Ferraresi

Defended by

Giuseppe Malara

Graduation Session 12/2024

Academic Year 2023/2024

ABSTRACT

With recent advancements in technology the modern landscape of job searching is being reshaped. This change affects a multitude of fields such as translation and linguistics, leaving industry professionals and recent graduates worried about the requirements needed to retain their position or enter the workforce. To investigate this issue, a study was carried out by Ferraresi et.al in 2021 as part of the UPSKILLS project, which sought to develop new curricula to address skills gaps for linguistics students. The study involved the creation of the UPSKILLS corpus, composed of online job postings for positions requiring skills related to linguistics and technology. A corpus-based analysis was then carried out to determine knowledge, skills and competences required of jobs at the crossroads between language and technology; as well as identifying the most salient tasks and responsibilities of these jobs. This thesis retains the same objective of the previously mentioned study and seeks to provide an update to it; both by providing standalone results and by comparing the two corpora. This is achieved through the creation of the WEBCTRL corpus, a carefully annotated corpus of jobs advertisements from various types of websites targeting the same kind of profile as the UPSKILLS corpus. Results show that the findings from the analysis of the UPSKILLS corpus remain valid, as the most salient skills and competencies (data and research skills, technical skills, language and linguistics disciplinary knowledge and communication, interpersonal and organizational skills) have not changed. The key difference is that technological requirements are being increasingly requested by employers, and more traditional linguistic requirements are on a downtrend. Artificial intelligence in particular represents a new requirement compared to the UPSKILLS corpus that holds particular importance in the WEBCTRL corpus. Future studies could employ more thorough corpus annotation and analysis and take into consideration mandatory and preferred requirements to best identify the key requirements needed for this specific job profile.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

1.1 The UPSKILLS project and scope of the thesis

The increasing integration of technology into every aspect of our lives has produced numerous changes, and its effects are starting to become visible in today's job market. The surge in technological requirements is not limited to programming or IT positions anymore, and is starting to affect other sectors like linguistics. Translators and linguists may be experiencing a revolution comparable to that of the adoption of Computer-Assisted Translation (CAT) tools, a staple of today's translators but not initially embraced without resistance. Similarly, the new wave of artificial intelligence (AI), large language models (LLM) and other related tools may be reshaping the profession.

Not only do the requirements for translators and linguists appear to be shifting, but new positions are also emerging, making a diverse skillset that integrates translation expertise, technological proficiency, and research capabilities a necessity. It is possible that this shift could result in a growing need for linguists who possess both traditional language skills and the ability to work with various digital tools and methodologies. This tendency has only become clearer in the last few years, and recent graduates or people currently enrolled in a linguistics-focused degree may feel uncertain about the specific skills and competencies required to thrive in the evolving job market. Their academic curricula may not fully address the technical and interdisciplinary knowledge now increasingly sought after by employers, leaving them potentially underprepared for roles that demand expertise beyond traditional linguistic training.

These concerns were initially addressed by the UPSKILLS project[1] between 2020 and 2023. UPSKILLS was an academic initiative that "sought to identify and tackle the gaps and mismatches in skills for linguistics and language students through the development of a new curriculum component and supporting materials to be embedded in existing programs of study". The project began on the assumption that tech giants such as Google, Amazon and Facebook were already working with language data and as such, the need for tech-oriented linguists was growing. However, some may feel that linguistics-related degrees might no longer offer them a good level of employability in light of these new requirements.

In order to design and produce the content necessary to achieve their end goal, the first action of the UPSKILLS project was to undertake a needs analysis to explore the status of the academic

[1] https://upskillsproject.eu/

offer in languages- and linguistics-related fields, the requirements of the job market in this area, and the gaps and overlaps of the two. The output of this research was the report "Competences, skills and tasks in today's jobs for linguists: Evidence from a corpus of job advertisements" by Ferraresi et.al in 2021.

The first objective of the report was to provide an overview of the knowledge, skills and competences graduates in language-related degrees or professionals might need, as well as of the typical tasks and responsibilities associated with these positions. To do so, a corpus of job advertising targeting these job profiles was created by accessing company websites, job websites and linguistics-focused websites; and was then annotated and analyzed. The three main categories that were analyzed were required qualifications (further divided into categories such as academic requirements or disciplinary concepts), job duties and job title. Four main types of skills required from employers emerged from the analysis: data and research skills, technical skills, language and linguistics knowledge, and communication and interpersonal skills. It must be specified that since the overall objective of the UPSKILLS project was to tackle the gap in skills in jobs spanning across languages and technology, standard profiles such as translator or reviser were excluded from the analysis.

The goal of the present work is to provide an update on the above analysis carried out by UPSKILLS and so retains the same overall objective: to examine jobs "at the crossroads between languages and linguistics, technology and research" (ibidem:17), and to identify the requirements and duties that these kind of jobs entail. To do so, a corpus of job advertisements was created much in the same vein of the UPSKILLS corpus, and methods were analyzed and updated where necessary while keeping comparability with the original corpus a priority. The end result of this thesis is the *Work and Employment Beyond Conventional Translation and Linguistics* corpus, or WEBCTRL corpus.

Before delving into the core of the study, sub-chapter 1.2 provides an overview of this work.

1.2 Thesis structure

This thesis is articulated in the following chapters:

Chapter 2 tackles concepts necessary to fully understand the work that will be carried out in later chapters by reviewing germane literature for each sub-chapter. The first sub-chapter will be dedicated to corpora, starting from the definition of corpus and briefly going over their various types before focusing on web corpora, which represent the focal point of this work. Then, data collection methods for corpus construction will be touched upon, ranging from manual and semi-automatic collection methods to automatic data collection and web scraping tools, all of which were considered for WEBCTRL's construction. The final subchapter initially handles job advertisements and their history, before delving into previous work in job ads for skill identification and language industry surveys and studies tailored to discover topical skills needed to thrive in the industry.

Chapter 3 explains the methodology used to build and annotate WEBCTRL. First, the text selection criteria for the preparatory phase of corpus building are outlined, and changes from the original UPSKILLS methodology are highlighted. Then, the corpus construction process is explained in detail, from initial attempts with various corpus building methods to the option that was ultimately chosen. A scheme is also provided to summarize corpus construction, and the chapter contains an overview of corpus annotation and the tools used to carry it out. Moreover, it also focuses on corpus description and methodology of analysis of the WEBCTRL corpus. Relevant metrics such as the size, composition, sections, sources of the WEBCTRL corpus are provided, and the pseudo-XML schema used to annotate the corpus is explained in detail. Comparisons with the UPSKILLS corpus are drawn, and differences or decisions behind any changes are specified. The same is done for corpus analysis methodology, broken down by object of study.

Chapter 4 presents results of the analysis carried out following the methodology outlined in the previous chapter. Results are first presented for the analysis carried out on the WEBCTRL corpus then. Since the UPSKILLS corpus has been re-analyzed following the same methodology, results are compared across the two corpora. The analysis is split into sections, and each section is complete with tables depending on the section. During comparison between WEBCTRL and UPSKILLS, tables putting results from the two corpora side-to-side are provided to highlight key differences and similarities.

Chapter 5 provides a summary of the thesis, draws conclusions from the results obtained and discusses future work.

2 Background

In this section previous research will be analyzed in the context of this thesis. Topics can be divided into three macro-categories: corpora, job advertisements, and language industry surveys and studies.

First, an introduction on corpora is given while the characteristics of a corpus are gradually delved into, then a brief overview of different types of corpora with a focus on web corpora is provided. Finally, a section on different data collection tools and methods for corpus construction is presented. Afterwards there will be an introduction on job advertisements and their history, followed by a literature review on the subject with a focus on studies for skill identification. Finally, the focus will be placed on language industry surveys and studies, in order to offer a different perspective on the questions of this thesis, namely, to identify the knowledge, skills and competences required for jobs at the crossroads between language and technology and the tasks and responsibilities that these positions entail.

2.1 Corpora and their design

A corpus is described by McEnery as "a well-organized collection of data, collected within the boundaries of a sampling frame designed to allow the exploration of […] certain linguistic feature[s]" (2003: 449). In light of this definition, it is important to note that corpora are no longer being used only to support hypotheses about a specific language feature or phenomenon in traditional linguistic studies; they are also employed in information extraction (Gundlapalli, 2015) or model training (Ma et al, 2023) in computational linguistics. Therefore it is important to focus on the substantial amount of human judgement that goes into selecting a collection of texts suited for a specific purpose.

Tognini-Bonelli (2001: 10) outlines three main points to take into consideration during corpus construction, namely "the authenticity of the texts included in the corpus, their representativeness and the sampling criteria used in the selection of the texts". In order for a text to be described as authentic, it must derive from genuine communication in either written or spoken form. If one were to create a corpus to study the language of native English speakers, samples from a textbook for English as a Second Language (ESL) learners would not be suited for such a corpus, and would stray from the notion of authenticity.

A corpus is representative when the inferences drawn from it can be extended to a larger sample of the population. Gray, Egbert and Biber (2017: 1) identify two types of representativeness in corpora: target domain representativeness, which "determines the generalizability of a corpus sample to a larger population of interest"; and linguistic representativeness, that is to say "the extent to which a corpus contains the full range of linguistic distributions that exist in the target domain".

Finally, sampling criteria are determined depending on the type of corpus one wants to build. Corpus construction begins with a clear delineation of its intended purpose and scope. It is essential to establish the overarching goals of the corpus, whether it is intended for linguistic research, language teaching, or specialized domain analysis for example. Additionally, the languages, genres, and time periods to be encompassed must be specified, ensuring that the corpus is tailored to the research objectives.

2.1.1 Types of corpora

While corpora can have a very large number of features, and some of them will be explored when referencing certain corpora, the main distinction that will be highlighted in this part of the thesis is the difference between general and specialized corpora.

The purpose of a general (or reference) corpus is to be representative of an entire language, as "it aims to be large enough to represent all the relevant varieties of a language, and the characteristic vocabulary […]" (Sinclair, 1996, Reference corpora, para. 1). It is therefore common for general corpora to include a considerable number of documents and a variety of text types, written or oral, making them ideal for linguistic comparisons with more specialized corpora.

An example of a very large and diverse general corpus is the Corpus of Contemporary American English (COCA). It contains more than one billion words of data from a wide range of genres such as newspapers, academic journals, movie subtitles, web blogs and more. Not only is the amount of data by genre balanced, but also each year ranging from 1990 to 2019 contains a similar number of words, making it a diachronic corpus. COCA was released in early 2008 and has grown from around 385 million words (Davies, 2009) to more than a billion in 2023. Because of these features, it can also be placed among monitor corpora, which are described by Sinclair as "gigantic, slowly changing stores of text" (1982:4).

Specialized corpora can be used to study a variety or type of language, for example political speeches or Shakespearean novels. They are deemed "specialized" not because of the level of specialization of the documents, but rather for "a number of specific pre-established criteria as a guide to selecting the types of texts to be included in the corpus" (Lopez-Mateo & Olmo-Cazevieille, 2015: 301).

As an example of a specialized corpus, I would like to examine a very different corpus from COCA: TWITTIRÒ (Cignarella et al., 2018). This is an Italian corpus containing tweets annotated for irony using a multi-layered annotation scheme. In comparison to COCA, this corpus is relatively small, totaling 1424 tweets and 28.387 tokens. While this corpus is part of a project containing also French and English corpora, it cannot be defined as a multilingual corpus since, in itself, it only contains tweets written in Italian. The authors mention how they could not retrieve the texts through a (semi) automatic method like they did for the other languages, and instead extracted tweets from already available corpora where the presence of irony is marked. Despite this, TWITTIRÒ can be called a web corpus, as its contents can be defined as naturally occurring instances of language on the internet.

2.1.2 Web corpora

Over the past decades the internet has grown to become a significant part of our lives and has changed them substantially, from mundane tasks to specialized interests. Of course, significant changes have been brought to the field of corpus linguistics as well, and the use of the web as a linguistic resource has become increasingly common. As the web increased in popularity, two approaches for the use of web data in corpus linguistics emerged, namely the "Web as Corpus" and "Web for Corpus" (De Schryver, 2002:267) approaches, which will serve as an introduction to this chapter, leading to more specific issues related to web corpora.

The main staple of the "Web as Corpus" approach is to treat the whole World Wide Web as an enormous corpus and to use, for example, Google's interface to access the contents of this "corpus". While it is true that the web contains a large amount of textual data, several objections have been raised to this approach over the course of the years. Sinclair (2005) mentions two crucial issues. The first is related to size, as all the resources used by researchers until then had been of a finite quantity. The web is ever-growing, and therefore we cannot know "how large" it is, making quantitative research not ideal if not outright impossible.

Sinclair also brings attention to the composition of the web, as when exploring results very little contextual information about the texts is provided. It is not simple to know whether a text on the internet has been edited or published on a whim, or if it was written by a native speaker and so on. This problem has become increasingly common and shows no signs of disappearing with the rise of Artificial Intelligence. As Robert McMillan from the Wall Street Journal (2023) points out, hundreds of fake news websites were found to be using AI to generate content, while instances of political misinformation and targeted message creation for hacking were also detected. In one case almost

2000 webpages were created and published in a few hours using AI to purposely steal away traffic from competitors and lead users to misleading content (Ward, 2023). This action can dilute the visibility of other relevant pages and capture traffic, as it affects Google's search engine optimization (SEO) and, viewing this from the Web as Corpus point of view, directly violates Tognini-Bonelli's (2001) principle of authenticity. If the web becomes rife with AI-generated articles, then its contents cannot be described as genuine communication anymore.

The Web for Corpus approach was popularized by Baroni and Bernardini (2004) with the release of their BootCaT toolkit (which will be analyzed in more detail in section 2.2.1) and has remained the most common method for web-based linguistic studies. Bernardini et al. (2006) referred to this approach as "Web as Corpus shop", meaning that the internet becomes a kind of store from which texts can be downloaded and added to offline corpora. Although the Web for Corpus approach has been widely adopted by the scientific community, the key issue brought against it is one of representativeness. Leech (2007:135) states that "the study of a corpus can stand proxy for the study of some entire language or variety of a language". This was achievable with more conventional corpora because the authors could better gauge the sample population, while it is difficult to have a good understanding when sampling texts from the web.

As has been seen before, the current method of building web corpora is to use the internet as a "corpus shop" and to search, download and extract texts from it. Other than data collection, which will be discussed in the next chapter, Paquot and Gries (2020) identify three important steps during the corpus building process for web corpora, namely boilerplate removal, document filtering and duplicate removal. The term "boilerplate" is often used to describe copyright notices, navigation menus, html code (if saving webpages as html) and other undesirable items that need to be removed from the texts. Since manually dealing with boilerplate is very time-consuming, the use of (semi) automatic tools such as jusText or PyRex, which incorporates many other useful corpus-related functions, is recommended.

Document filtering is often applied when using (semi) automatic methods for data collection, as a way to include in the corpus only documents with certain characteristics. The most common are language or size filters, so as to include only texts in a certain language, which is key when building a monolingual corpus, or to include only texts with a certain number of tokens and exclude documents that might be too long or too short and skew the corpus. Stop-word filtering or other forms of spam filtering are also common. The relevance of the duplicate removal step is self-explanatory, as it is not ideal to have multiple copies of the same text present in the corpus. For this reason it represents a crucial step in the corpus building phase of almost all corpora. While compilers of smaller corpora could perform duplicate removal manually, larger corpora are almost always forced to use tools such as Onion (ONe Instance Only) (Pomikálek, 2011) which have been integrated into Python and have already been used to perform this step on large textual collections at Masaryk university in Brno.

2.2 Data collection methods for corpus construction

2.2.1 Manual (and semi-automatic) data-collection

While the web is no doubt a very valuable source of textual materials, manually collecting texts from the web can be very time-consuming, since tasks such as text selection, acquisition and sometimes transcription (when dealing with spoken corpora) will fall entirely on the person creating the corpus. A keen eye is required to distinguish reliable sources from unreliable ones, to deal with boilerplate, download errors, and typos if transcribing text. Nevertheless, constructing a corpus manually rather than through web crawling can lead to a higher degree of accuracy in text selection, a higher quality of the sources used, and can represent a valuable alternative where more automatic methods might be considered superfluous or difficult to implement.

Tessari (2017) created a comparable Italian-German corpus on the 2014-2016 European economic crisis. Its main intended application was the creation of a terminological and phraseological glossary as a resource for interpreting students. The Italian corpus totals around 270.000 tokens, and is composed of manually selected parts of already transcribed parliamentary sessions handpicked from the official website of the Italian Chamber of Deputies. Every transcription from the two-year period of the economic crisis was examined and either added to the corpus or discarded, since results from the internal search engine of the website were deemed not precise enough. The German corpus totals a similar number of tokens and was built using the same method employed in the creation of the Italian one, using parts of transcribed parliamentary sessions from the German Bundestag. The search engine of the official website of the German federal parliament too was found lacking and perusal of transcriptions was carried out by hand, rendering the data collection part of this paper completely manual.

As an alternative to manual corpus construction, semi-automatic methods can also be used. These methods are often employed when completely automatic tools such as web scrapers are not utilized for data collection, either due to their perceived complexity or because the corpus size does not necessitate their use.

Ever since the paper *BootCaT: Bootstrapping Corpora and Terms from the Web* by Baroni and Bernardini was published in 2004, the titular toolkit has become the de-facto standard in semi-automatic corpus construction. BootCaT requires a small set of seeds (terms) from the target domain as input, which are then randomly combined to form tuples and each tuple is used as a Google query string. Webpages containing the tuples are downloaded and converted to text, new seeds are then extracted from the corpus and combined with the previous ones for another round of queries. This process can be repeated as many times as necessary, and each feature of the BootCaT toolkit can be used individually for other tasks.

For example, Giampieri (2021) built an English corpus of contracts drafted in Italy and Poland to analyze in-context terms that may have an unclear meaning. In order to do so she queried Google for *"terms and conditions of" site:.it* (and *site:.pl* for Polish) and locally downloaded the first 10 results pages. The *local queries* mode of BootCaT was then used to retrieve the documents and convert them to a textual format.

Although BootCaT and its features can still be successfully employed for corpus construction purposes, changes to the structure of the internet and to search engines' application programming interfaces (APIs) represent a constantly evolving roadblock. Furthermore, Barbaresi (2015) finds several shortcomings in the BootCaT method: search engines present unknown biases due to their increasingly commercial nature, their APIs may prove too expensive or limiting, webpages load content (such as images, advertisements etc.) from a growing number of external sources, creating multiple points of failure for semi-automatic corpus collection.

Due to these shortcomings, one may be inclined to turn towards more automatic methods for the data collection phase of corpus construction, such as web scraping.

2.2.2 Automatic Data Collection and Web Scraping tools

To delve deeper into automatic data collection for corpus construction, one must get acquainted with the concepts of web crawling and web scraping. Even though these two terms are often used somewhat interchangeably, Barcaroli et al. (2016: 27) state that a web crawler "systematically browses the Web starting from an Internet address […] and some predefined conditions". Web crawlers are used by search engines as Google or Bing for indexing purposes, as they can associate pages with certain words to the query made by the user (Panta, 2015). A web scraper instead "takes Web resources (documents, images, etc.), and engages a process for extracting data from those resources […]" (Barcaroli et al. 2016: 27).

The main differences between crawling and scraping software can therefore be found in their underlying purposes and scope of action: while crawlers act mainly on the internet to index and download webpages in bulk, scrapers act on a set range of pages with the objective of extracting specific data for further processing or analysis.

Understanding the structure of web pages is crucial for effective web scraping. Websites are built using Hypertext Markup Language (HTML), which defines the structure and layout of a web page. The Document Object Model (DOM) represents the hierarchical structure of HTML elements. By navigating and manipulating the DOM, web scrapers can pinpoint and extract the desired content. To scrape a website, web scrapers send HTTP requests to the server, which responds with the requested web page's content.

However, websites are continuously adding measures to protect themselves against excessive scraping. A prevalent issue is rate limiting, a mechanism employed by websites to restrict the frequency and speed of incoming requests, acting as a deterrent against excessive scraping. Websites may also restrict all forms of scraping, pointing users towards using their APIs in order to access their data. Since anti-scraping techniques are becoming increasingly common, it is getting harder for researchers to acquire data through completely automatic methods. Datadome, a prominent bot protection and anti-scraping solution, lists websites such as TripAdvisor, Rakuten, LeParisien and Patreon among their clients.

Nevertheless, web-scraping tools have evolved to find methods to bypass anti-scraping measures, or one could choose to focus on websites that are similar to the target one to avoid dealing with such limitations and still acquire relevant content. The most commonly used tools for web scraping are programming libraries (a collection of pre-written code) that can be accessed from the command line interface (CLI) to perform various tasks, and standalone software, which aims to simplify web scraping through a user interface.

Lotfi et al. (2021) have written an extensive literature review on the most advanced web-scraping techniques and tools. Their study also provides an overview of their application in various fields, general implementation methods, and of the different types of technology that can be employed. The paper focuses entirely on programming libraries and CLI-based tools. While an explanation for this choice is not provided, one could speculate that one of the reasons is the higher degree of customization, integration, and flexibility these types of tools can provide. The paper concludes that, while most web scrapers are similar, Scrapy provides better speed, power, and scalability over competitors such as BeautifulSoup or Selenium.

Matta et al. (2022) instead conduct a comparative study of various GUI-based web-scraping tools. While the study does not define a clear-cut best solution like the previous paper, it puts together a list of pros and cons of the five tools it sets out to analyze, which could prove very useful for researchers with different degrees of coding experience. For example, a linguist who would like to avoid coding may use a tool such as import.io: an easy-to-use web-scraper with an intuitive UI that requires no coding, but with an expensive subscription model where each sub-page costs credit. If one's programming skills are more advanced, Dexi.io has a modular interface fit to build custom tools for different websites with a high degree of integration at the cost of a steeper learning curve.

2.3 Job Advertisements and corpus analysis

2.3.1 A brief summary of job ads

It will be remembered that the text type that makes the object of this study is that of job advertisements which, while maintaining their defining characteristics, have also undergone important changes since the advent of the web. Job advertisements, also referred to as "job ads" or "job postings" are "documents acknowledged through public media for the company or the organization to find the right talents to fill in vacant positions" (Fu, 2012: 399-400). The job posting represents a long-standing text genre: its emergence in newspapers is traced back to the mid-1800s by Redman and Matthews (1992: 29) who, at the time of publishing, still identified newspapers as the "most common medium for recruitment advertising".

While job ads were predominantly present in newspapers in the past, they were subject to space limitations due to a word-based advertising cost, so much so that Bruthiaux (1996:90) described their style as concise and standardized, and classified them as instances of "linguistic simplicity". In the same paper, Bruthiaux (ibid:126) finds five main components of the standard job ad, namely:

1. *The target* (or job description)
2. *The recruiter* (or information about the company/placing of the ad)
3. *The requirements* (qualifications and/or experience)
4. *The reward* (salary and benefits)
5. *Contact segment*

The transition to digital job advertisements has introduced a new modality and semiotic elements, transforming the traditional paper-based format into one that fits a dynamic, interactive platform replete with buttons, images, and various other features. However, despite these changes, the five components identified by Bruthiaux appear essentially unchanged. Garzone (2018) finds that, other than the addition of a “job summary” section, the main differences in the aforementioned components lie in their order and realization. For example, information about the hiring company is much more detailed due to online job ads being free to publish or subject to a flat rate, while the *reward* component seems to have lost prominence. The employers seem to have developed a tendency to only mention the benefits in passing (e.g. “Salary 30.000€ + benefits”) or to omit them completely. This is surprising, as previous research has shown that this information influences initial application decisions. In a study by Van Hoye and Lievens (2005), it is stated that job ads containing this type of information are even capable of reducing the effects of negative publicity about a specific company, to an extent.

2.3.2 Job ads research and related studies

Job adverts have extensively been used as a data source in research in various fields such as content analysis, corpus linguistics, information research and more. Harper (2012) offers an overview of the main purposes of papers in the recruitment advertising literature. The main goals defined in the study are skills identification, that is examining “the changing nature of skills required in the workplace” (Harper, 2012:29), curricula development, tracking changes in the job environment or market, and assessing the situation of job positions that are hard to categorize. Harper does not take into consideration linguistic or sociocultural analysis, but it is important to note that job advertisements have also been investigated in this field.

Gaucher et al. (2011) conducted five different studies to prove the existence of gendered wording in job advertisements and how it sustains gender inequality. First, they sampled job advertisements for male and female dominated jobs, used already available lists of masculine and feminine words to identify them in their corpus, and finally conducted a 2x2 analysis of variance (ANOVA) test. The results seemed to confirm the hypothesis that job ads within male-dominated areas contained more masculine wording and vice versa. In their other experiments, psychology students were exposed to the previously mentioned job advertisements and were asked to rate their appeal and anticipated belongingness using a Likert scale. Results found that, within job ads containing more masculine wording, participants perceived more men within these occupations and women found these jobs less appealing.

Before delving into more linguistically adjacent studies, it is important to note that all of the works that have been mentioned and that will be mentioned below analyze job advertisements that were posted online. It can be argued that this is due to the internet’s spike in popularity in the last 15-

20 years (not only as a job searching resource, of course), and newspapers and newspaper ads becoming less prominent than they once were.

The Boston Consulting Group and the Recruit Works Institute (2017) conducted and published one of the largest global surveys on job seeking, sampling more than 13.000 people from 13 countries. This survey includes an investigation on job searching channels, where it emerged that internet job sites are the most common method of job searches across all 13 countries, with paper-based media coming in at third, behind referrals from colleagues, family, or friends. The closest gap between the two can be found in India, where 39% of respondents name internet job sites as their job searching channel of choice, against 21% who choose paper-based media. The largest gap can be found in Russia, where only 5% of respondents chose paper-based media as their preferred method of job searching, against 44% of the votes going to internet job sites. The results seem to be consistent even across age groups. Online job sites are still the most popular option across all ages, while referrals are shown to gain a bit more popularity among older people, with paper-based media hovering around 9% across all age groups.

Mang (2012) finds that job changers who looked for their new job online are better matched to their newly found job than people who found it through newspapers or other channels. The estimation model for calculating job matching quality includes individual-level covariates such as gender, age, migration status, education, number of job changes in the last seven years, and a dummy indicating whether the person was unemployed during the last year. Karacsony et al. (2020) focus on the job searching habits of Generation Z, that is people born between 1995 and 2010, by conducting a survey with more than 230 responders. Not only are online job sites shown to be more popular than newspapers among Generation Z, but surprisingly social media emerge as the most popular job searching method. While the fact that LinkedIn, one of the most known job searching websites in the world, is considered as a social media in the analysis can be a point of contention, 63% of respondents stated that they received the most job offers from Facebook, while other prominent social media such as Instagram or TikTok are used as a way to obtain information about a specific company.

Even though corpus studies that analyze job advertisements are present in previous literature, it seems that the most common methodology in this field is content analysis. Kutter and Katner (2012:7) describe content analysis as:

> a standardised hermeneutic procedure of text interpretation in the course of which the individual analyst assigns abstract categories to propositional contents [...] that occur in passages of the analysed texts. The categories do not correspond to the 'observable' (linguistic) characteristics of the text, but to hypotheses derived from social theory about the social and political setting in which these texts occur [...].

Instead, corpus linguistics is a broader discipline which uses a set of more mildly delineated methods to study discourse; and while procedures such as the use of concordancers are well established, other aspects are not (McEnery and Hardie, 2012). Despite corpus linguistics being seemingly less defined than content analysis, several papers have tried to integrate the two approaches by creating "corpus-based content analysis". For instance, Kutter and Katner (ibidem) suggest a heavier focus on the linguistic elements of the text, a feature they feel had little recognition in content analysis. In order to perform quantitative analysis of large text samples, they developed a content analytical procedure, stating that the steps are the same for corpus linguistics and content analysis, but the way they are achieved differs. In concept specification, the corpus approach would focus on the terms deemed more relevant to the concept being investigated, while content analysis would "make out the content dimensions of the concept" (ibidem:15). In the second step, content identification, a corpus-based method would investigate lexical items through the use of concordance, wordlists, keyness and collocate analysis; through conventional content analysis one would instead conduct "an inductive pilot analysis in order to figure out which content variants typically represent the selected content dimensions" (ibidem:16), and would start the construction of a codebook. In the concept assignment step, the occurrence of relevant items identified in the previous step is investigated in corpus methodology, while content analysis would start categorizing the content dimensions of the concept according to the previously established codebook. The final step, quantitative analysis, is practically the same for both disciplines, but it is carried out on different items (lexical items/documents vs. codes).

Tarat et al. (2021) also try to integrate the two disciplines. The main difference between these two studies stands in the fact that the latter uses corpus methods to create a framework for content analysis, while Kutter and Katner seemed keener on creating a new method that combines the two disciplines. To illustrate their method, Tarat et al. (ibidem) create and analyze a corpus of LGBTQ+-related research articles published on SAGE Journals, a platform for peer-reviewed academic journals. After downloading and converting the texts, these are categorized by time period and uploaded to a corpus analysis platform where keyword lists are generated and keywords with a chosen keyness value are selected. At this point, content analysis methodology takes over, as codes and categories are assigned to the keywords and the "Discussion & Conclusion" section of the articles are used as data to be coded, and finally content analysis is performed on the coded data. For example, in this experiment words such as "equality, marriage, adoption" are found to be highly representative of LGBTQ+ in the category of civil rights and law.

While it could be argued that content analysis differs from corpus methods, job advertisements are collected in content analysis studies and form a kind of corpus, as exemplified in Do Vale et al. (2018). In this study, literature review was performed and job ads were collected and analyzed in order to investigate the competences of project managers. A systematic review was carried out across 88 articles, and highlighted competences such as leadership, planning and organization. The 449 job ads collected show that most vacancies require a university degree and previous experience in project management, while the ability to work with project management software formed a requirement in only 20.4% of the jobs.

2.3.3 Job ads for skill identification

In this sub-chapter various content analysis and corpus-based studies will be reviewed. As seen before, the two methodologies share a substantial common ground, and are grouped together in order to focus on the objective of these studies, which is either the same or resembles very closely that of this thesis. The studies that will be reviewed aim to identify the knowledge, skills, and competences of one or more professions, and/or to identify the most prominent companies in a specific sector through the collection and analysis of job advertisements.

Lipovac and Babac (2021) employed content analysis on a corpus of job advertisements to identify and provide a broader overview of employability skills required among different professions and countries. The study collected 16.000 job ads from the website Indeed from 16 different cities, with most of the cities being in the USA and several in the United Kingdom, plus Dublin and Hong Kong. While 1000 postings have been collected for each city, the specific method through which these job postings have been collected is not detailed in the paper. The key parts of job advertisements being analyzed are job title, company, job description and salary. As for the first category, results highlight similar trends shared among the US, the UK and Ireland, with the most vacancies being found in administration and service providing activities, while projections for a high amount of jobs available in the healthcare, science, technology, and engineering sectors are not reflected in the data. Hong Kong differs in that the most common vacancies are in consultancy, finance, and administration. The most important data extracted from the company part of the ads is that Hong Kong is the only city where private companies are the most common type of employer. From the job description analysis it emerges that employers require a wide range of skills even for low-skilled occupations, and that social and communicative skills are the most sought after. An interesting fact is that employers tend to put the emphasis on company branding rather than the skills they require from potential employees. Salaries were briefly touched upon in a single paragraph, with the main point being that job advertisements tend not to mention them very often.

Sodhi and Son (2009) also use content analysis to infer required skills in operational research (O.R.)[2] by analyzing more than a thousand online job advertisements. It must be noted that, although data collection took place in 2005, it is stated that the ads were posted in a timeframe that ranges from 1999 to 2005. It is also important to mention that the amount of postings per year is not defined and the study does not adopt a diachronic approach, therefore there is the possibility that some of the skills were no longer as important already at the time of publication. Since the target for this study are graduates with O.R. degrees, the researchers collected ads from two popular job sites and an academic and industry specialized online journal, finding relevant content by keyword search. They then created a dictionary, variables and categories by listing all words and phrases and dropping the ones that appeared in less than 2% of the ads, and then grouped them in related categories while including the shorter phrasing for similar sentences. Jobs were then analyzed using the frequency of ads by category, cross-tabulation, and spearman rank-order correlation. The results highlight four skills that characterize jobs in the operational research field, namely modeling, statistics, programming, and general analytical skills. They are deemed as such because they feature in the top six requirements regardless of sector, function, or degree type.

Bowker (2002) provides an investigation of the terminologist profession in Canada through corpus methods. She collected more than 300 job ads in order to provide insights on terminology in Canada, and touched upon many topics such as types of job, requirements, duties and more. It must be noted that most of the ads in this study, probably due to the time of publication, come from newspapers and job bulletin boards, even though a small part of online ads is included as well. The author states that, at the time, terminology was almost at its infancy compared to other professions, and therefore included postings that do not explicitly require a terminologist but for which terminology experience might be a required or a nice-to-have skill. After compiling the corpus in electronic format, the analysis was then carried out with corpus tools (WordSmith Tools and MonoConc Pro). Corpus analysis revealed how only 13 ads out of the 308 collected in total were explicitly directed towards dedicated terminologists, and another 53 ads mentioned knowledge of terminology as a requirement, while most of the postings were directed towards professional

[2] Operational Research can be described as the practice of converting real, complex operational problems into mathematical ones, solving them using various methods, and then transforming the mathematical solutions back into real-world terms. Source: https://www.lancaster.ac.uk/~blackb/whatisOR.html

translators. These 66 job advertisements were then further analyzed, with health, finance, technology and business surfacing as the domains where terminology was most needed. As far as qualifications were concerned, most of the openings required the candidate to hold a degree, 50% specified that it should be a degree in translation, while 8% required a degree in linguistics. Previous experience was also a prevalent requirement, but employers often failed to mention the specific years required, and when they did they usually required 2 years of experience. Computer skills were the most sought after with 72% of the ads mentioning them as a requirement, closely followed by research skills, domain knowledge and being a “team player”. Bowker (2002:15) makes an interesting statement when analyzing the employment conditions: while most of the ads advertise full-time, permanent positions, much of the work in language industries is offered on a freelance basis, which ends up being underrepresented in this analysis.

2.4 Language Industry Surveys and Studies

It must be understood that the language industry plays a pivotal role in this thesis, since all of the job advertisements that have been collected will effectively belong to this sector. This sub-chapter will focus on the most relevant results of surveys on the language industry, and will review studies that have the same topic. The objective of touching upon the language industry is twofold: on the one hand it provides an overview on the state of the industry this work is interested in; on the other, surveys can be considered a different method of obtaining the same answers this study is seeking to provide, and can offer significant insight.

The European Language Industry Survey (ELIS) is an annual initiative funded by the European Commission that focuses on the trends, expectations, and concerns of language service companies (LSC), independent contractors and other professionals in the industry. The 2023 edition (reporting on 2022 data) of the survey highlights how LSC and independent language professionals are drifting apart, both in terms of performance and market sentiment. Smaller language service companies and freelancers show significantly poorer economic results, with only 5% of companies reporting an increase in profitability. Independent professionals cite a lack of fair remuneration, work-life balance issues, lack of clients and instability as their main causes of concern. LSCs do not share the same impressions, with work from home being on the uprise for the employees and investment scores, which reflect the intention to create or acquire other companies, surging back to pre-COVID levels.

Although human translation was by far the most common type of service in 2022, it lost ground by 10% compared to 2021, showing a negative trend. Technology is instead on the uprise, with MT services nearly doubling in 2022 compared to 2021, although LSCs, independent contractors, and academia, feature this sector among both the positive and the negative trends. While artificial intelligence was viewed as a promising avenue in the previous years, AI is now perceived as a challenge, with ELIS linking this trend to the negative coverage of ChatGPT in press and forums. Language service companies are shown to be very worried about data security requirements, and the level of concern about increased workload combined with shrinking staff is alarming.

In the "Entering the language industry" part of the survey, results show that around a quarter of young professionals (that is, people who have been working for less than two years) have created their own company and therefore have become independent contractors, while another quarter has joined their current company after an internship. The rest of the votes are evenly split among entering via job advertisements, peer references, spontaneous applications and recruitment agencies. The vast majority of young professionals is on a permanent contract. ELIS also focuses on skill gaps perceived by academia and by the language industry, with the two differing significantly. For example, translation technology was the primary concern for language service companies in previous editions of the survey, but this year LSCs feel there have been substantial improvements by graduates, while academia still sees a gap between expectations and performance that continues to widen. The same could be said for digital literacy and information gathering/processing: academia feels these skills need to be improved, while LSCs feel that candidates have honed their skills in these regards. While a section specifically on skills required by the industry is not present in this survey, one could speculate that digital skills are more relevant than ever, considering their relevance in the previously mentioned section.

Unfortunately, academia and training institutes continue to be plagued by the same issues detected in 2022, namely severe budget and administrative constraints and a waning interest in pursuing a career in the language industry. It is interesting to note that literary translation features as the second most common core subject in training programs while not being cited in any other part of the documents, and services such as audio description, for which LSCs have reported an increase in demand, are a core part of very few degrees and are often offered as an optional course. This is also a major cause of concern among students, who cite the difficulty of combining studies and a profession as a major challenge, a worry which ranks behind only the uncertainty of the future and time pressure.

The language industry is ever-evolving and has been at the center of academic research too. Ferraresi et al. (2021), the main work this thesis is based on and part of the UPSKILLS project, conduct a corpus-driven analysis on a corpus of job ads of the skills, competences, tasks and responsibilities present in postings for graduates with language related degrees in positions in the middle of translation and technology. The advertisements for the corpus were manually collected from company websites, employment platforms and other sources, and contain almost 200 items. The text selection criteria were decided on the basis of previous UPSKILLS research, and involved excluding jobs for which a degree in a STEM field was required, or jobs involving "almost exclusively" (ibidem: 5) translation or revision tasks. The corpus was then annotated in an XML-like format with various sections based on the metadata (e.g., the "jobtitle" section or the "requiredqualifications" section) and analyzed through corpus methods on the NoSketch Engine platform. The results highlight the importance of computational linguistics in these jobs, as it is the second most frequent collocate for the expression "degree in", and it is mentioned in 25.4% of the job ads collected (44.7% if including Natural Language Processing, which is counted as a separate discipline). It also emerges that data and research are very sought-after skills, as they feature both as skills that companies expect the candidate to possess, but also as (part of) activities they are expected to carry out in their role. A notable aspect of this study is the fact that university degrees constitute a requirement in less than half the samples, and sometimes they represent only a preferred qualification rather than a required one. Researchers suggest that this phenomenon can be linked to various problems affecting language and linguistic degrees.

On this topic, a study by Maulan et al. (2023) aims to determine if the curricula design and the courses of language degrees fulfill the requirements, both in terms of skills and knowledge taught, of the industry. It does so by involving 13 key players from the language industry among owners, executives and freelancers with a minimum of ten years of experience, and by having a three hour interview with the participants. Before the interview, the experts were provided with the program's academic structure and course information of the University Teknologi MARA in Malaysia, the institution concerned in this work. Experts were also involved in assessing the students' final semester project and evaluating their knowledge and skills. In the results, communication skills are cited as the most important and in-demand quality a candidate can possess, and it is concerning that experts found Malaysian graduates to be lacking in this field. This type of skills was deemed essential when dealing with clients or replying to feedback, while negotiation and persuasion skills are vital in situations such as determining a professional's rates, and personal and interpersonal skills were described in a similar fashion. The need for computer-based skills, such as word processing, data entry, and knowledge of the most used language service tools was also voiced, along with marketing skills to market products and establish a social media presence. Finally, students need to master research skills, described in this study as “the ability to search for information on products, services, and clients, doing need-analysis, [...] finding the right techniques for market survey and [...] assist in understanding copyright issues [...]” (ibidem: 769). Overall, industry experts agreed that the curriculum design meets the industry needs, but they suggested strengthening the existing courses and preparing the students for the workplace.

2.5 Summing Up

Previous literature shows that content analysis and corpus analysis are the two predominant methods for skill identification in job advertisements, especially since advertisements have started moving from paper media to the digital space. Advances in corpus construction techniques have enabled more automatic methods to build larger, more comprehensive corpora; although semi-automatic methods can still be employed to focus on accuracy while still retaining a respectable corpus size. From industry surveys, it is possible to gather that communication skills remain very important for recent graduates aiming for roles inside the language industry. Technological and research skills also hold substantial relevance, although the perception of the candidates’ proficiency level differs among LSCs and academia.

The following chapter outlines a corpus-based methodology adapted from the UPSKILLS framework, with specific modifications to capture the skill requirements essential for recent language graduates pursuing careers at the intersection of translation and technology.

3 Methodology

In this chapter, the methods adopted to conduct this work are reviewed, and details are provided for each of the stages. First, the text selection criteria are presented and explained. Then, each step of the corpus construction phase is described, including preliminary methods that were later scrapped from the project (section 3.1.2). Afterwards, the methodology and tools employed to annotate and analyze the final corpus are outlined and expanded upon in section 3.1.4. In section 3.2 the final WEBCTRL corpus is described. Relevant metrics and statistics are provided, and comparisons with UPSKILLS are made. The pseudo-XML schema used for corpus annotation is broken down and discussed, and the various parts and decisions behind any additions or removals are explained. The analysis carried out on both the WEBCTRL corpus and the UPSKILLS corpus is described. The employed methodology is discussed and explained, together with any modification to methods previously employed in the UPSKILLS project. Tools settings and research parameters are also defined and discussed for reproducibility of the study in the future.

3.1 Corpus construction and description

3.1.1 Text selection criteria

The criteria for text selection were based on those already chosen for the initial UPSKILLS needs analysis, since the broad objective of this thesis consists in providing an update of the previous work, adopting similar methods and a comparative perspective. The job profile that is targeted is a position, mainly aimed at recent graduates, involving "language or linguistics-related tasks requiring digital and/or research skills" (Ferraresi et al., 2021:5). It can be argued that some degree of arbitrariness is involved in the definition of text selection criteria for this task, as there is not a specific, traditional job position such as "salesman" being analyzed. That is why the criteria applied in the original UPSKILLS report have been examined, modified, and changed in some cases; while trying to preserve comparability as much as possible. The reasons for each criterion will be explained in detail, while similarities and differences with the UPSKILLS report will be discussed and highlighted.

Job ads were researched starting from the sources included in the original UPSKILLS report (full list available in Annex 1); they served as a starting point for research, which later expanded to other websites. It also must be noted that, as in the original UPSKILLS report, only advertisements in English were considered and no attempt was made to select or restrict the location where the positions are offered. The original reasoning (companies having an increasingly global nature and jobs being more and more open to remote and hybrid positions) still stands. As an addition to these requirements, only job ads posted in the last year were considered, whereas the only constraint in the original paper was that the advertisement had to be available online. This change was made to offer a more focused and updated analysis of the trends and needs in this sector. Advertisements requiring a degree in a STEM field, such as computer science, were excluded, as were those involving only "a) content creation tasks (e.g. writing/editorial jobs), or b) translation and revision tasks" (ibidem:5). These requirements are kept unchanged from previous work, as the target audience is recent graduates with a linguistics background, but the positions must require some degree of digital skills. For the same reason, job posts mentioning several years of experience as a mandatory requirement were also excluded.

In the original UPSKILLS work, academic jobs were actively excluded from search, and jobs from institutions and (trans)national bodies were not explicitly targeted. In this work, both types of jobs were included in the search. The reasoning behind this choice was that the requirements from both the academic and the public sector can offer interesting insight, should they differ from those of the private sector. Unfortunately, only one job ad from (trans)national bodies was found during the search, and it had to be excluded as it did not fit the text selection criteria. A small number of job advertisements from the academic sector did instead fulfill the requirements and were added to the corpus. Finally, no change was made to the decision of only keeping one version of the same position offered to candidates with different language competencies (e.g., "Linguist – Italian" and "Linguist – English").

3.1.2 Corpus construction

Corpus Construction for WEBCTRL was an iterative process that underwent several changes before reaching its final state. The goal was to find the right amount of automation without sacrificing the quality that comes from manual construction of corpora, and many avenues have been explored in this regard. Initially, several web-scraping solutions were considered in order to obtain a bigger amount of data. Initial solutions included using tools such as Octoparse[3] and Scrapy[4] to download web-pages in bulk, and building an ad-hoc solution for this study was also considered. Unfortunately, as previously discussed in section 2.2.2, websites have significantly increased their countermeasures against web-scraping, and pursuing this direction proved to be rather challenging. The most popular job-searching platforms such as LinkedIn or Indeed turned out to be the ones with more advanced measures in place to prevent data extraction, and attempts to obtain information from these sources did not yield acceptable results.

[3] https://www.octoparse.com/

[4] https://scrapy.org/

An attempt to employ AI-based solutions was also made. Microsoft's Copilot, an AI-tool with GPT technology and access to the internet[5], was used in an attempt to retrieve job advertisements from various websites such as LinkedIn or Amazon Jobs. Even though results looked promising initially, the plan to go forward with this solution was scrapped when it was noted that the tool was generating data when unable to access the target webpage. This made the risk of analyzing inauthentic data non-negligible, and plans shifted towards a more controlled approach.

The first step in the corpus construction process consisted in checking all of the sources present in the final UPSKILLS corpus to make sure the websites were still online, and could yield useful results. Since the final UPSKILLS corpus follows a naming scheme composed of *websitename_textnumber*, one text for each prefix was opened and the website was inserted into a list. Once this process was finished, each site in the list was opened in a browser window to check its contents. The majority of the websites were unchanged and were used for this thesis's research; however, there were a few that were either offline or did not return useful results anymore. The most important example is the Career Linguist website[6], of which the "Job postings" section was explored in the original UPSKILLS report. However, as of the writing of this work, the website does not offer the aforementioned section anymore. In the end, this process resulted in a list of 13 websites which could potentially contain relevant job postings.

The next step was to create queries To inspect the contents of the websites contained in the list. This query construction step is rooted, once again, in the original UPSKILLS report. The paper used the keywords "linguist", "linguistics", "data" and "language specialist", while specifically excluding "NLP". This was done because the use of this keyword often returned results for jobs that were too tech-oriented, and relevant results could be found regardless, using the keyword "linguist". Therefore, the Initial query to be submitted to Google was *"linguist"* | *"linguistics"* | *"data"* | *"language specialist" site:specific.website*. Therefore at least one of the words between quotes had to be in the webpage, the "|" character is equal to the "OR" Google operator, meaning that one or more of the words between quotes must be in the result, and the "site:" operator specifies which website the query will be conducted on. This query was applied to all of the websites obtained in the previous step.

[5] https://www.microsoft.com/en-gb/store/b/copilotpro

[6] https://careerlinguist.com/

In order to download and extract the Google results, two browser extensions, namely linkclump[7] and SingleFile[8], and the BootCaT toolkit were used. First, the query was inserted into the Google search engine, then, once the results page was displayed, the "continuous scrolling" setting was enabled. This setting makes it so that by scrolling at the end of the page, other results, if present, will keep appearing until no more results satisfy the research parameters. This allowed me to load all the links to the query results into a single webpage. At this point, the linkclump browser extension allowed all of the results to be opened in a new window at once. Each window was loaded into memory, and then the SingleFile browser extension allowed the opened webpages to be downloaded simultaneously.

In this way, 140 job advertisements were retrieved. In order to gauge how efficient, useful, and precise the employed method was until then, manual inspection and filtering of the files was performed. Each job posting was opened and requirements were checked. It transpired that 47 texts could not be included in the corpus. In these texts the position in question was too technical or not technical enough, required a degree in another field or required too many years of experience. A further 59 texts were on the edge of fulfilling the requirements, but needed further examination before inclusion in the corpus. Most of the uncertainty was about the degree requirements and how technical the task at hand was. This left only 34 texts whose inclusion in the corpus was certain. The primary conclusion to be drawn was that the implementation of (semi)automated methods is difficult in a task with fuzzy criteria such as this one, hence, a more manual method was employed. The previously mentioned 59 texts were ultimately not included in the corpus. One might argue that this level of uncertainty about the technical know-how and degree types that are required is a core feature of the texts under analysis, and therefore not a reason for excluding them, but this decision was taken to limit the number of jobs with requirements that are unlikely to be fulfilled by a recent graduate.

In the second round of text collection, queries similar to the initial one were performed on Google, but the list of websites was expanded, and a search was performed specifically on job-seeking platforms instead of Google, when possible. Websites of technological or translation companies were added based on the author's knowledge and based on the most popular companies mentioned in LinkedIn postings. In an attempt to find new sources, queries without a "site:" operator were performed. The word "job" was used in place of the "site:" operator (e.g., *"language specialist"* | *"linguist" "job"*). Queries also evolved in an iterative way. For example, if a job as a "solution analyst" was found to satisfy the requirements, this job title was entered as a query in a subsequent

[7] https://github.com/benblack86/linkclump

[8] https://github.com/gildas-lormeau/SingleFile

round of searches. LinkedIn was by far the platform with the most results, however, its job search function does not allow the use of operators, and often resulted in unrelated advertisements much more often than related ones. To try to get around this obstacle, queries exclusively for LinkedIn were devised and implemented. Since all job postings on the platform start with "*linkedin.com/jobs/view*" in the URL, queries were written with keywords and the "site:" operator that included the previous URL. This formula was reused for promising websites with limited searching capabilities. Figure 1 summarizes the queries that were ultimately used for corpus construction and provides an overview of the methodology employed in query building.

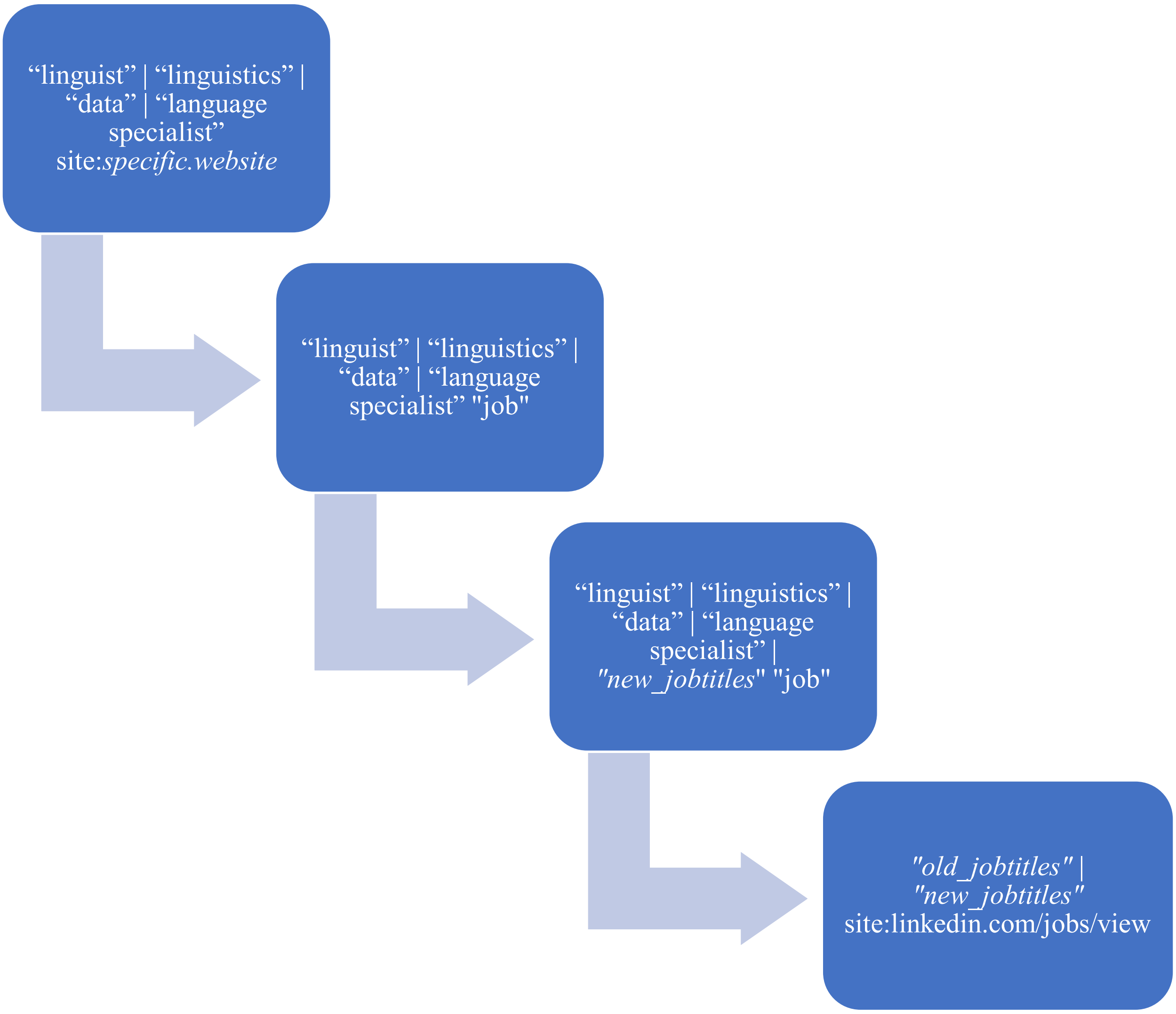

"linguist" | "linguistics" | "data" | "language specialist" site:*specific.website*
"linguist" | "linguistics" | "data" | "language specialist" "job"
"linguist" | "linguistics" | "data" | "language specialist" | *"new_jobtitles"* "job"
"old_jobtitles" | *"new_jobtitles"* site:linkedin.com/jobs/view

Figure 1: Evolution of queries for WEBCTRL corpus construction

Each result was manually opened and inspected before downloading. Once requirements were ascertained to be fulfilled, the webpage was downloaded in .mhtml format. This file format allows one to save a complete webpage, including all embedded resources. This file format was chosen mainly to account for the “see more” functionality on LinkedIn or other job advertisement platforms, where some content is hidden until this button is pressed. If webpages were saved in .html, information could potentially be lost, therefore, the .mhtml file extension was chosen. Unfortunately, this choice came with the drawback of .mhtml not being a popular file format and having very limited compatibility with other programs. For example, BooTCaT could not extract text from .mhtml files, and manual copy-pasting had to be employed to convert all texts to .txt files.

3.1.3 Corpus annotation

After completing the text collection step, each job ad was annotated following a pseudo-XML schema, much in the same vein as the original UPSKILLS corpus. The main objective of this annotation schema is to be able to store and represent the contextual metadata and structural information of each text, while allowing for thorough and efficient analysis of the corpus.

CTRL+1	<section name="jobtitle">…<\section>
CTRL+2	<section name="keyinfo">…<\section>
CTRL+3	<section name="jobdesc">…<\section>
CTRL+4	<section name="jobfunctions">…<\section>
CTRL+5	<section name="requiredqualifications">…<\section>
CTRL+6	<section name="about">…<\section>
CTRL+7	<section name="benefits">…<\section>
CTRL+8	<section name="company">…<\section>
CTRL+9	Wraps the text in a structure of which the user can manually input the name

Table 1: SublimeText shortcuts for corpus annotation

In order to fully annotate of each .txt file, the Sublime Text[9] text editor was used. Sublime Text is a cross-platform text editor designed for coding purposes. It provides features such as syntax highlighting, auto-completion, and a package ecosystem for additional functionalities and customization. This software enabled partial automation of an otherwise heavily manual annotation process. Specifically, the Snippet functionality was used to create a quick way of wrapping text in the XML-like structure present in the original UPSKILLS corpus. Sublime Text Snippets are programmable smart templates that can automatically insert text and adapt it to context. This solution

[9] https://www.sublimetext.com/

meant that text just had to be selected, and then a shortcut would automatically wrap the text in the selected structure, instead of having to manually type each individual structure. Table 1 presents all of the shortcuts and the respective sections they add, while the full snippets can be found in Annex 2.

3.1.4 Corpus description

The WEBCTRL corpus is a corpus of job advertisements mostly targeted at recent graduates in language-related degrees. The positions included in this corpus require a combination of language, research and digital or technological skills. Only ads in English published in the last year were preserved. Table 2 contains information on corpus size and composition, while figure 2 displays the various text sources.

Tokens	72,498
Number of texts	111
Average text length (+ Standard Deviation)	588.9 (233.6)
Number of companies	73
Average number of texts per company (+ Standard Deviation)	1.5 (1.4)

Table 2: Size and composition of the WEBCTRL corpus

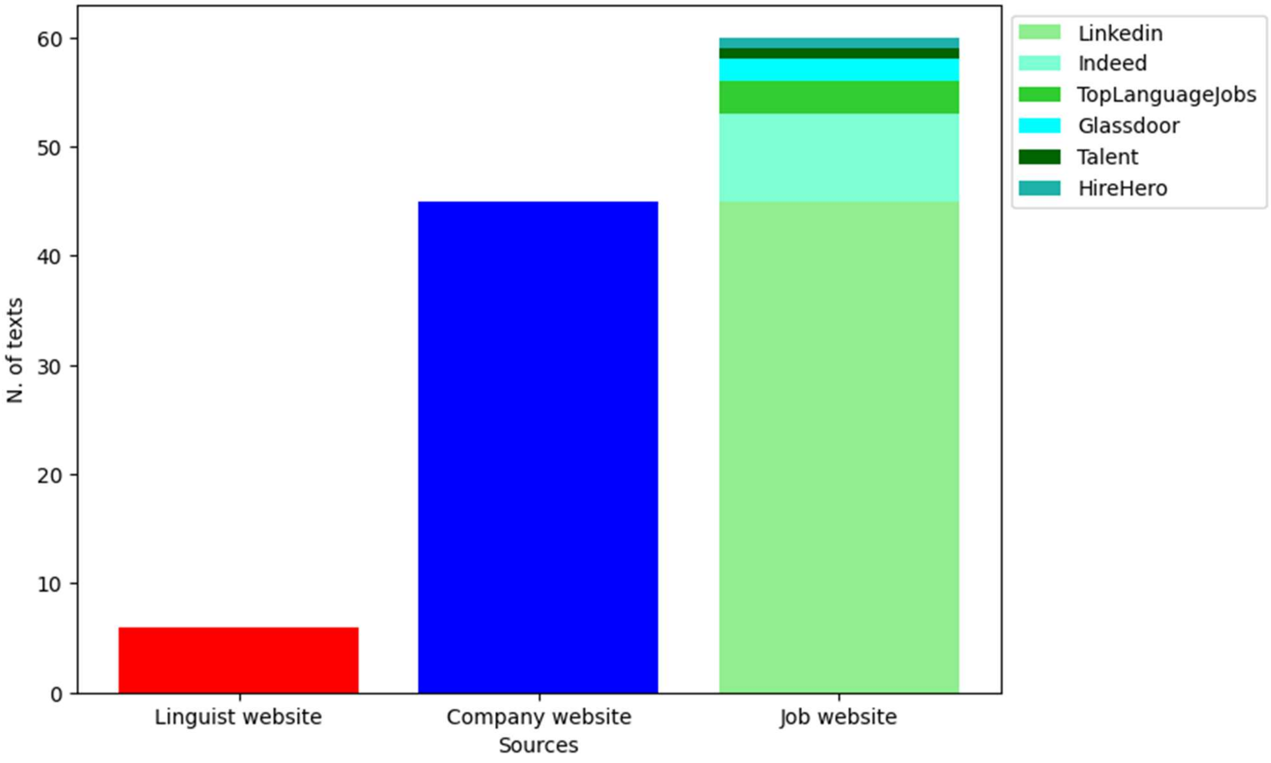


Figure 2: Sources of WEBCTRL texts

WEBCTRL is relatively small in size, for various reasons. Firstly, using automatic and semi-automatic text collection methods yielded less than satisfactory results, and the adoption of fully manual methods was necessary. Pivoting to a completely manual method slowed the text collection process significantly, and led to the creation of a smaller corpus, albeit one that includes only well-curated results. Text sources are divided into three main categories: Linguist websites, company websites, and job websites. Linguist websites suffered a sharp decline compared to the UPSKILLS corpus, and make up around 5% of the corpus, with the rest of the job ads coming from either company websites or job websites. LinkedIn is the most prominent source, making up around 40% of the corpus. This may point towards LinkedIn being the most used job website by companies offering this type of job. It may be speculated that this is the case because LinkedIn is the only website among the examined ones that allows users, employers and employees alike, to establish a social media presence and expand their reach. This unique dual role of job board and social media platform may have had a hand in making LinkedIn the most popular job website, which in turn may have simplified the hiring process just by virtue of the very wide adoption rate of both recruiters and candidates. In comparison, texts from LinkedIn make up only 20% of the UPSKILLS corpus, which offers a more balanced distribution. Texts from job websites were still prominent and are represented in 40% of the corpus, with the remaining 60% being almost evenly split between linguist websites and company websites.

It will be remembered that comparability with the UPSKILLS corpus is an important concern of this work, and has influenced the corpus design and analysis method in multiple ways. The general approach has been to replicate the UPSKILLS methodology while trying to improve on it where possible. For example, an attempt was made to create a more automated corpus construction method, even though in the end a manual method was employed due to unreliable results. An assisted annotation process was successfully created as mentioned in section 3.1.3, whereas the UPSKILLS procedure did not focus specifically on this step. Existing structural divisions were kept mostly the same to allow for easier comparison, but some were added to facilitate more granular analysis in future studies.

To allow a high level of granularity in the analysis, and to keep comparability with the UPSKILLS corpus, texts were annotated following a pseudo-XML schema. At the start of all the texts, there is a header that contains specific information on text origin and metadata, while sections wrap parts of each text. Table 3 and 4 show the full information contained in the header and the sections that are present in the texts.

<text id="Amazon001"
source="Company website"
company="Amazon"
job_name="Content Developer">
	id = File ID, the name of the file.
	source = Type of text source (company website / linguist website / job website)
	company = Name of the company publishing the offer
	job_name = The job title offered as specified in each ad

Table 3: Example of text header in the WEBCTRL corpus

<section_name="jobtitle">...</section>	Name of the job
<section_name="keyinfo">...</section>	Summary of the job ad
<section_name="jobdesc">...</section>	Core part of the job ad
<section_name="jobfunctions">...</section>	Job functions, what the candidate will need to do in the job
<section_name="mandatoryqualifications">...</section>	Qualifications that the candidate needs to have in order to apply
<section_name="preferredqualifications">...</section>	Qualifications that the company views positively for the job
<section_name="requiredqualifications">...</section>	All the qualifications mentioned in the ad; includes both mandatory and preferred qualifications
<section_name="about">...</section>	Information about the company publishing the ad
<section_name="benefits">...</section>	Pay, perks, training opportunities and benefits
<section_name="company">...</section>	Company publishing the offer

Table 4: Pseudo-XML schema used in the WEBCTRL corpus for sections

In the text header included with each job ad, the element *source* specifies the type of text source (job website, linguist website and company website). In the case of job websites, the name of the website was included for ease of access, since only six websites were consulted, while Linguist List is the only linguist website included in the corpus. As for company websites, too many were visited to include in the *source* element, and this information is available in the *company* element. The job_name element includes the name of the position as published by the company. However, there were occurrences where additional information was provided directly in the job title, which was omitted when redundant. For example, a job with the title “Project Manager – Italian” would be tagged as *<job_title=”Project Manager”>, so as to increase its generalizability*.

As for the sections present in each text, the *benefits* section includes information on both the salary and any additional benefits. This is because there were not enough texts where both were mentioned to warrant including a separate *salary* section. Certain advertisements did not group information about benefits in a single or in consecutive paragraphs; in such cases, multiple *benefits* sections were included to wrap pieces of text with the relevant information.

Three sections are included to describe academic or professional requirements a candidate might need to apply for the job in question: *mandatoryqualifications, preferredqualifications*, and *requiredqualifications*. Mandatory qualifications are defined as those the employer overtly mentions as “required” or “needed”: without these types of qualifications it is assumable that the candidate will not get the job. Preferred qualifications are described as “additional” or “nice to have”: a candidate without this type of qualifications can still be employed. Required qualifications include both aforementioned categories. This enables the possibility of a detailed analysis of the different types of job requisites. It also facilitates the creation of a more general overview of every type of qualifications desired by employers. Moreover, this allows for a comparative analysis with the homonymous section of the UPSKILLS corpus. The WEBCTRL corpus arguably offers a finer level of detail for analysis with the addition of three separate categories for requirements in the annotation scheme, namely *requiredqualifications, mandatoryqualifications* and *preferredqualifications*.

The *url* and *file* elements are not present in the WEBCTRL corpus. The former was not included due to the volatile nature of job advertisements, as they are often promptly removed after a candidate is hired, while the second element contains a direct link to each file in the UPSKILLS corpus, and it is not included in the WEBCTRL corpus.

The distribution of text sources differs between the two corpora. While the UPSKILLS corpus has a balanced distribution among texts coming from job websites, company websites and linguists websites, the WEBCTRL corpus is mainly composed of texts from LinkedIn and company websites. There is a sharp decline of jobs coming from linguists’ websites, which is mostly due to the absence of texts from the CareerLinguist website, which no longer hosts job offers. Table 5 presents corpus statistics for the two corpora, while table 6 offers a comparison between their text sources.

	WEBCTRL	UPSKILLS
Tokens	72,498	107,421
Number of texts	111	197
Average text length (+ Standard Deviation)	588.9 (233.6)	544.1 (242.7)
Number of companies	73	112
Average number of texts per company (+ Standard Deviation)	1.5 (1.4)	1.8 (2.1)

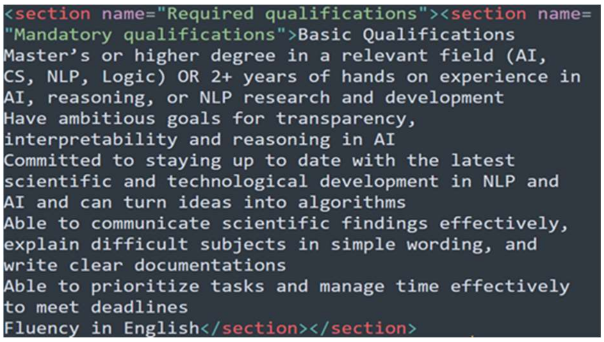
Table 5: Comparison of size and composition of the WEBCTRL corpus with the UPSKILLS corpus

Text source	WEBCTRL	UPSKILLS	% difference
Job platform - LinkedIn	45 (40.54%)	38 (19.29%)	+21.25%
Job platform - Other	15 (13.51%)	48 (24.37%)	-10.86%
Company Website	45 (40.54%)	58 (29.44%)	+11.10%
Linguist Website	6 (5.41%)	53 (26.90%)	-21.49%
Total	111 (100%)	197 (100%)	

Table 6: Comparison of number of texts per source between WEBCTRL and UPSKILLS

Finally, the texts were uploaded to the Sketch Engine[10] platform for analysis, using the default settings for the corpus upload process. The only exception is that nested structures were enabled (a non-default setting). This choice was made due to the formatting of the requirements sections: The *requiredqualifications* section contains occurrences of *mandatoryqualifications* and/or *preferredqualifications*, this is an example of what is referred to as a "nested structure". By default, Sketch Engine does not allow for nested structures, as the application is expecting the first section to end with a closing tag before the next section begins, and would therefore not detect all instances with two "sub-sections". Therefore, the choice was made to enable nested structures by editing the corpus configuration file and changing the value for the nested attribute for the section class from 0 to 1. Modifying the nested attribute changes this behavior and allows for refined searches within all the aforementioned sections. Figure 3 provides an example of the text formatting of the various requirements sections.

[10] https://app.sketchengine.eu/



```
<section name="Required qualifications"><section name=
"Mandatory qualifications">Basic Qualifications
Master's or higher degree in a relevant field (AI,
CS, NLP, Logic) OR 2+ years of hands on experience in
AI, reasoning, or NLP research and development
Have ambitious goals for transparency,
interpretability and reasoning in AI
Committed to staying up to date with the latest
scientific and technological development in NLP and
AI and can turn ideas into algorithms
Able to communicate scientific findings effectively,
explain difficult subjects in simple wording, and
write clear documentations
Able to prioritize tasks and manage time effectively
to meet deadlines
Fluency in English</section></section>
```

Figure 3: Example of nesting in the "qualifications" sections

3.2 Corpus analysis

The corpus data was analyzed entirely on the Sketch Engine platform. The methodology used to carry out the analysis follows the same corpus-driven, bottom-up principles of the UPSKILLS corpus analysis (Ferraresi et al.,) unless stated otherwise. Since one of the main objectives of this work is to gauge if the status of jobs “at the crossroads between languages and linguistics, technology and research” (ibidem) has undergone changes, and if so, to assess the entity of these changes; corpus analysis has been carried out from scratch on both corpora. This ensures that methodologies remain consistent across studies, and allows for an easier comparison. Section 5.3 compares results among the two corpora and highlights important differences.

For each of the three main sections (Required qualifications, Job functions, Job title) the most common lemmas, 2-4 grams, or noun phrases containing a noun pre-modified by another noun or an adjective, were identified. For each section, the list of the most common lemmas was generated via the wordlist function of Sketch Engine, and NTLK’s list of English stop words[11] was employed to ensure that the search ignored irrelevant items. Lists of the most common 2-4 grams were generated using the n-grams function of Sketch Engine, while were identified using the query *[tag="NN.*|JJ.*"][tag="NN.*"]* in each section of the corpus.

The top results from each operation were then manually analyzed and inspected through the use of collocate generation and concordance analysis. The goal of this process was to find significant and easily identifiable categories in each section, much in the vein of the exploratory approach used in the UPSKILLS project, and in line with previous literature in section 2.3.2. After identifying a category, ad-hoc queries were devised in order to target that specific category, as the methods used for preliminary analysis were not meant for more precise, focus analysis.

[11] https://gist.github.com/sebleier/554280

For instance, as will be expanded upon in the next chapters, one of the top bigrams in the Required Qualifications section was “bachelor’s degree”. This prompted the creation of the “formal education requirements” category, which was targeted to answer real-world questions such as “what is the most requested degree for the job profile of this analysis?”. The following chapters expand on the analysis methodology employed in each of the three main sections of the corpus, and briefly mention categories, when found, to elaborate on the queries employed in that specific category.

3.2.1 Required Qualifications

After exploratory analysis, two main categories emerged: formal education requirements and experience and knowledge. This last category is very broad, and takes on many different meanings. Therefore, as will be explained in section 3.2.1.1, the decision was made to employ different queries to try and break down “experience and knowledge” into neater categories. For example, experience could refer to “experience in conflict resolution”, which would be a soft skill, or “experience in Microsoft Excel”, which would constitute a more technical requirement. This was made in order to gather a better idea of what the experience and knowledge required by employers in this field really entails.

In the UPSKILLS project, it is mentioned that the qualifications found in the “Required qualifications” section are not necessarily mandatory to apply for the job, and can also contain “preferred qualifications”, as in requirements that provide an additional asset to the employer and increment the chance of the candidate to be selected among similarly skilled applicants. In order to prioritize comparability with the UPSKILLS corpus, the “Required qualifications” section was analyzed, which includes both “mandatory” and “preferred” requirements, although separate analysis of “Mandatory qualifications” and “Preferred qualifications” is possible in the WEBCTRL corpus. The next step after identifying the two main categories was to focus on each one separately. They are illustrated in detail in the following subchapters, methods enlisted for each category are also described.

3.2.1.1 Formal education requirements

For this category, an attempt was made to identify the most requested degree levels (bachelor’s, master’s, etc.) and the fields to which the degrees belong. The first query was carried out with the aid of the wordlist tool in Sketch Engine: 13 keywords were employed, and the search targeted the “Required qualifications” section of the corpus. The keywords included items such as “bachelor”, “bachelors” and “bachelor’s”, in an attempt to include different spellings or synonyms to catch all occurrences for a certain degree level. The full list of keywords is available in annex 1. It is important to note that the first query makes no distinction between more linguistic-oriented degrees and STEM-oriented or technology-oriented degrees, as job ads with out-of-scope requirements were already excluded during the corpus construction phase.

To find the most requested degree fields, the collocates of the lemmas “degree in” were analyzed in a span of +10 and a minimum frequency in corpus of 3. Initially a span of +5 was utilized, similarly to the method in the UPSKILLS report. However, after manual concordance inspection, it was noted that ads often post an extensive list of degrees that usually span more than five tokens. Therefore, in order to keep the search as accurate as possible, results were manually analyzed to make sure one excludes possible irrelevant items (e.g. “degree in [...] or relevant translation experience”, where “translation” would be picked up as a degree type).

3.2.1.2 Experience and knowledge

This subcategory encompasses various requirements and, therefore, multiple queries were employed. “Experience” can refer to the candidate having been employed in a similar role and having performed tasks that will be required, or having being previously hired for a number of years for the same position as that of the ad. To address this issue, the query *[lemma="year"] []{0,4} "experience"* was used in this section of the corpus, and collocates in a span of -4 were analyzed.

Experience can also refer to the ensemble of tools, abilities, and concepts relating to the field of work at hand that the employer expects the candidate to be familiar with. It could be said that “experience”, “knowledge” and “understanding” of “skills” and “abilities” are crucial for a potential candidate. Therefore, collocate analysis in a span of ±5 was performed separately for each lemma previously mentioned in quotes, concordances for each collocate were checked and interesting data was recorded in the “additional information” column for each collocate.

These queries were useful in retrieving tools and soft skills, but a different, more focused query was put in place to specifically look for disciplinary concepts. A search for the most common noun+noun or adjective+noun phrases was carried out, while excluding soft skills. This query was devised on the basis of preliminary analysis, where it was observed that academic results were found often enough to warrant the use of the query while filtering out the occasional irrelevant result.

3.2.2 Job functions

For this section of the corpus, the initial exploratory analysis was not very informative. Therefore, the decision was made to analyze the top 50 lemmas and carry out in-depth collocate and context analysis to extract more complete information on the job tasks a candidate is expected to carry out.

Collocates were looked for in a span of ± 3 with a minimum frequency in corpus of 5. Not all 50 lemmas were indicative of the job functions required, or were directly attached to another lemma and its context (for example, *feedback* is part of the *training* collocates, but the candidate is expected to give *training* and provide *feedback.* Therefore *feedback* was excluded as it was considered redundant). After carrying out analysis of all 50 lemmas, only 28 were deemed relevant.

3.2.3 Job title

As noted by Ferraresi et al. (2021), “the Job title section usually consists of a single phrase, which may be idiosyncratic to the job post being analyzed, or to the company publishing them” (ibidem:15). It would therefore not be advisable to carry out the same kind of analysis on this text type. Therefore, in this paper, the same methods used in UPSKILLS are directly replicated: a word cloud based on a frequency list from the “Job title” section was drawn while excluding phrases appearing in less than 5% of the corpus. Additionally, an analysis of the most frequent 2-to-3 grams was also carried out and top occurrences were categorized in accordance with the most relevant keyword of the title (for example, the job title “Italian Freelance Linguist would be categorized as “Linguist”).

4 Results

4.1 Introduction

This section presents results of the analysis carried out following the methodology outlined in chapter 4. First, results from analysis carried out on the WEBCTRL corpus will be presented. Each main document section of the corpus (required qualifications, job functions, job title) will have its dedicated sub-chapter in which results for each query will be presented and expanded upon. After all the results from the WEBCTRL corpus are presented, results will be compared with the analysis carried out on the UPSKILLS corpus, to find similarities and differences between the two corpora and to identify any changes concerning the job profile targeted in this work. As mentioned in section 3.2.1, exploratory analysis revealed a range of requirements that was divided into two main categories: formal education requirements, and experience and knowledge. The following sections illustrate the results of the analysis in these two categories.

4.2 Results – WEBCTRL

This sub-chapter will focus on the results of the analysis carried out on the WEBCTRL corpus, while 4.3 will tackle comparison of results between WEBCTRL and the UPSKILLS corpus.

4.2.1 Required qualifications – Formal education requirements

Table 7 shows the number of documents and percentage in corpus related to each degree level in the WEBCTRL corpus and illustrates how many documents mention each degree level and their percentage in the corpus (for example, a bachelor's degree is present in 41 documents out of 111, meaning it is present in 37% of the documents). It must be noted that some jobs may require a BA or an MA, meaning that a BA would be enough to satisfy the requirements. No documents explicitly stating that no degree is required were found during analysis.

Results show that a bachelor's degree is the most requested degree level for this job profile, as it is mentioned as a requirement in slightly over a third of job advertisements. A master's degree represents the second most requested academic qualification, and is present in 18% of job ads. Finally, a PhD is requested in only 11 documents (10% of the corpus), as can be seen in table 7, around three times less requested than a bachelor's degree.

Degree level	Number of documents	Percentage in corpus
Bachelor's degree	41	37%
Master's degree	20	18%
PhD	11	10%

Table 7: Degree frequency in WEBCTRL corpus

After analyzing the most requested degree types, the focus was shifted to degree fields. It must be specified that most of the ads list a number of degrees a candidate must possess, meaning that having one among all the listed degrees would fulfill the requisites. Table 8 shows that linguistics is the most requested degree in the corpus, appearing in around one document out of five. Although linguistics is present in more than double the number of documents than all the other degrees, the rest of the table is predominantly STEM-oriented. Computer science and computational linguistics rank second and third, and are present in almost the same number of documents: 10 and 9 respectively. Speech Science, Information Science and Applied Science account for a combined 4% frequency in the corpus, along with Mathematics. While it may be considered surprising, a mathematics course should provide a strong foundation in subjects such as statistics and probability, which would be of use for jobs that delve into NLP and more tech-oriented tasks. Translation and Literature are present in 3 documents each and represent degrees in the humanities, while Information Systems and Information Science combined are present in 3% of the documents in the corpus. Finally, Language Technologies is present in only 2 documents (2%) across the corpus, perhaps due to it being a relatively new degree, or employers preferring to focus on either the linguistic or technological requirements of the candidate.

“Degree in”	Number of documents	Percentage in corpus
Linguistics (computational linguistics removed)	24	21%
Computer science	10	9%
Computational linguistics	9	8%
Speech science / information science / applied science	4	4%
Mathematics	4	4%
Translation	3	3%
Literature	3	3%
Information science / Information systems	3	3%
Language technologies	2	2%
TOTAL	**62**	

Table 8:Degrees mentioned twice or more in the Required qualifications section of the WEBCTRL corpus

4.2.2 Required qualifications – Experience and Knowledge

The analysis of this vast category first started by focusing on the years of experience companies require in the job advertisements they post. It must be noted that while these years of experience might be required, they might not necessarily be mandatory. Results point towards potential employers asking for 2-3 years of previous experience most of the times, as these requirements appear in respectively 15% and 11% of the corpus. One year of experience is required 6 times in the WEBCTRL corpus, that is as many times as five years of experience, while no job ad asks for more than five years of experience. This could point towards junior positions being the most common companies need to fill, or towards job titles or duties of the job being relatively new. It must also be reminded that job posts mentioning several years of experience as a mandatory requirement were excluded, and this might have been a factor in these results.

Years of experience	Number of documents	Percentage in corpus
1	6	5.41%
2	16	14.41%
3	12	10.81%
4	3	2.70%
5	6	5.41%
TOTAL	**43**	**38.74%**

Table 9: Years of experience required in WEBCTRL corpus

The collocation analysis of the lemma “experience” and of related lemmas (knowledge, understanding, skills, and abilities) revealed that requirements in this section can be divided into two main categories: *Translation and linguistics* and *Technical expertise*. Requirements that did not fit either category were put in the *Miscellaneous* category. This is a departure from the original UPSKILLS project, where the categories were *Language competences*, *Data, tools and techniques*, *Academic disciplines* and *Other*. This decision was made following preliminary analysis, which showed that the categories originally emerging from UPSKILLS did not fit the results of WEBCTRL. Moreover, as mentioned in section disciplinary concepts are being explicitly targeted with the noun or adjective + noun query so that no information is lost.

The analysis of the lemma “experience” will be presented first due to the large amount of results found, while the analysis of other lemmas will be presented as one, even if analysis was carried out separately for each lemma. The tables are ordered by lemma first, where present, then by category, and finally by percentage in corpus.

Table 10 shows results for the lemma “experience”. It can be seen that most requirements fit in the *Technical expertise* category. Within this category, the most common requirement pertains to working with various programs or programming languages, and the ability to analyze or annotate data. Programming itself is also among the top results, with Python being the only programming language to be explicitly mentioned as a requirement. Most of the requirements in this category can be seen as "general", with lemmas such as "tools", "develop" and "systems". Among more specific requirements, knowledge of AI, and generative AI in particular, is one of the most notable results, with knowledge of the Linux operating system and of the version control tool GIT[12] for programming also ranking highly.

Perhaps unsurprisingly, top results in the *Translation and linguistics* category include lemmas such as "linguistics" and "language". It is important to note, however, that language did not refer only to the knowledge of multiple or foreign languages, but occasionally also referred to language models and programming languages, making this collocate more tech-related than initially imagined. "Translation" and "localization" each appear only in 10% of the documents. Perhaps this is due to the text selection criteria applied during corpus construction trying to exclude traditional translation jobs, or it may point towards the job profile under analysis becoming more detached from traditional translation tasks and gravitating more towards more technological requirements. "Text" is another collocate that sometimes pointed towards tech competencies when used as part of "text classification".

Finally, results in the *Miscellaneous* category indicate that managing skills, in particular project management skills, may prove useful for this job profile. Communication skills, research skills and industry knowledge are also frequent requirements.

[12] https://git-scm.com/

Field	Collocate	Number of documents (Occurrences)	% in corpus	Contextual information
Translation and linguistics	Linguistics	19 (27)	17,12%	
Translation and linguistics	Language	18 (34)	16,22%	Models/multiple/foreign/programming
Translation and linguistics	Localization	11 (21)	9,91%	
Translation and linguistics	Translation	11 (17)	9,91%	
Translation and linguistics	Text	9 (11)	8,11%	Data/classification
Translation and linguistics	Ontologies	4 (9)	3,60%	
Technical expertise	Working with/in/on	40 (51)	36,04%	Data(ML/language/text/speech), vendors
Technical expertise	Data	25 (36)	22,52%	Analyze/annotate/types of data
Technical expertise	Using	16 (30)	14,41%	Various programs/programming languages
Technical expertise	Environment	15 (17)	13,51%	Scholarly/connected/related to skill
Technical expertise	Programming	14 (16)	12,61%	As in programming languages
Technical expertise	AI	13 (16)	11,71%	Generative / AI in general
Technical expertise	Python	13 (15)	11,71%	
Technical expertise	Machine	11 (11)	9,91%	Learning
Technical expertise	Software	10 (15)	9,01%	Developing / types of
Technical expertise	Analysis	9 (11)	8,11%	Data
Technical expertise	Tools	8 (10)	7,21%	Localization/types of
Technical expertise	SQL	8 (8)	7,21%	
Technical expertise	Develop	16 (21)	7,21%	Software/products
Technical expertise	Systems	7 (11)	6,31%	AI/types of
Technical expertise	Computational	7 (9)	6,31%	Linguistics
Technical expertise	Linux	7 (8)	6,31%	
Technical expertise	GIT	6 (8)	5,41%	
Technical expertise	Technical	6 (8)	5,41%	Skill/fields
Miscellaneous	Management	16 (25)	14,41%	Project/product/account/product
Miscellaneous	Project	12 (16)	10,81%	Management/coordination
Miscellaneous	Communication	10 (11)	9,01%	Skills
Miscellaneous	Industry	7 (14)	6,31%	
Miscellaneous	Research	5 (9)	4,50%	Processes/market/industry

Table 10: Collocates of "experience" in the "Required qualifications" section of the WEBCTRL corpus categorized

Collocation analysis for "Knowledge" and "Understanding" confirms the trend of technical requirements being more present in the corpus, as they appear in 48 documents, while the category "Translation and Linguistics" is present in 41 documents. Moreover, it is important to note that while the collocate "language" is classified in the aforementioned category, most of the instances of this collocate refer to programming language, with only three hits referring to language aspects.

As can be seen in table 11, the analysis of these lemmas reveals a different set of requirements, especially in the *Translation and linguistics* category. In this category, a need emerges for the knowledge or understanding of linguistic principles (with "design" referring to a specific job duty such as designing a curriculum or prompt design principles), structural aspects of languages, grammar and semantics, all with a very similar percentage in corpus of around 4%. Collocates in the *Technical* *expertise* category are mostly the same as those found in the analysis of the lemma “experience”, while this analysis also highlights the knowledge or understanding of more specific tools such as Microsoft Office, SQL (a database-oriented language used to manage data), regular expressions and scripting.

Field	Collocate	Number of Documents (Occurrences)	% in corpus	Additional Information
Translation and linguistics	Language	11 (12)	9,91%	Only 3 occurrences refer to foreign languages or language aspects, the other are programming languages
Translation and linguistics	Principles	5 (5)	4,50%	linguistic and design
Translation and linguistics	English	5 (5)	4,50%	
Translation and linguistics	Syntax	4 (5)	3,60%	
Translation and linguistics	Structural aspects of languages	4 (4)	3,60%	exact phrasing was found all times
Translation and linguistics	Grammar	4 (4)	3,60%	
Translation and linguistics	Linguistics	4 (4)	3,60%	
Translation and linguistics	Ontologies	3 (3)	2,70%	
Translation and linguistics	Translation	3 (3)	2,70%	(1) translation software
Translation and linguistics	Semantics	2 (3)	1,80%	
Technical expertise	Data	8 (8)	7,21%	mining, science languages, governance, analysis
Technical expertise	Systems	6 (7)	5,41%	database management/ git
Technical expertise	Scripting	6 (6)	5,41%	
Technical expertise	Python	5 (5)	4,50%	
Technical expertise	Regular expressions	4 (4)	3,60%	
Technical expertise	Natural	4 (4)	3,60%	3 language processing / 1 natural languages
Technical expertise	Programming language	3 (4)	2,70%	
Technical expertise	Tools	3 (4)	2,70%	technical, seo, statistical
Technical expertise	Microsoft office	3 (3)	2,70%	
Technical expertise	SQL	3 (3)	2,70%	
Technical expertise	Machine learning	3 (3)	2,70%	

Table 11: Collocates of "knowledge" and "understanding" in the "Required qualifications" section of the WEBCTRL corpus categorized

The last analysis that analyzes specific lemmas targeted “skills” and “abilities”. This analysis revealed soft skills that may have been underrepresented in previous analysis, while most results in both categories have already been found elsewhere.

From this analysis it emerges that over half of the jobs in the WEBCTRL corpus explicitly require some form of communication skills. Written communication skills are requested in over a quarter of the documents, while verbal communication is present in 17% of the corpus. In this case, the “Miscellaneous” category encompasses all types of soft skills. Interpersonal skills are the most common requirement in this category at around 12%, closely followed by problem-solving skills, management skills and attention to detail. It is also interesting to note that the ability to work independently and collaboratively while also being able to work under pressure is a requirement in more than 22% of the corpus, pointing towards a certain degree of versatility and adaptability being a need in this job profile.

The divergence in results between the first three lemmas (“experience”, “knowledge”, and “understanding”) and the last two (“skill” and “ability”) is noteworthy. While it could be argued that the theme of both queries remains the same, and linguistics and technology remain at the forefront of the analysis, it also seems that the first three lemmas returned results concerning concepts, topics and tools, while the last two lemmas seem to point more towards soft skills.

An indication of this phenomenon can be found in the fact that there are only 7 results that could not be categorized in the two main categories in the analysis of the lemma “experience”, and there were no results placed in the *Miscellaneous* category. Meanwhile, there are 14 results in the *Miscellaneous* category in the analysis of “skills” and “abilities”.

Field	Lemma	Collocate	Number of documents (Occurrences)	% in corpus	Additional Information
Translation and linguistics	Skills	English	18 (21)	16,22%	
Translation and linguistics	Skills	Language	11 (14)	9,91%	Mainly foreign
Translation and linguistics	Skills	Writing	9 (16)	8,11%	Various types of writing like technical, coding, content, business level English etc.
Technical expertise	Skills	Analytical	12 (14)	10,81%	
Technical expertise	Skills	Python	8 (10)	7,21%	
Technical expertise	Skills	Computer	6 (6)	5,41%	
Technical expertise	Skills	Programming	6 (6)	5,41%	
Technical expertise	Skills	Research	5 (10)	4,50%	
Miscellaneous	Skills	Communication	61 (76)	54,95%	
Miscellaneous	Skills	Written	31 (38)	27,93%	Mostly communication
Miscellaneous	Skills	Verbal	19 (25)	17,12%	Mostly communication
Miscellaneous	Skills	Problem-solving	21 (24)	18,92%	
Miscellaneous	Skills	Interpersonal	14 (17)	12,61%	
Miscellaneous	Skills	Management	13 (18)	11,71%	Project/time/people management
Miscellaneous	Skills	Detail	13 (13)	11,71%	As in attention to detail
Miscellaneous	Skills	Organizational	12 (12)	10,81%	
Miscellaneous	Skills	(Critical) thinking	6 (8)	5,41%	
Miscellaneous	Abilities	Work	25 (28)	22,52%	independently/ collaboratively/remotely / under pressure
Miscellaneous	Abilities	Manage	11 (11)	9,91%	multiple priorities
Miscellaneous	Abilities	Prioritize	8 (8)	7,21%	tasks
Miscellaneous	Abilities	Learn	5 (6)	4,50%	tools/software
Miscellaneous	Abilities	Multitask	4 (5)	3,60%	
Technical expertise	Abilities	Data	5 (5)	4,50%	use/analyze

Table 12: Collocates of "skills" and "abilities" in the "Required qualifications" section of the WEBCTRL corpus categorized

Finally, a query targeting noun+noun or adjective+noun phrases was carried out in order to find more disciplinary concepts. As previously mentioned in section 3.2.1.2, it was discovered during analysis that Sketch Engine misrepresented the number of occurrences of these phrases. Therefore, occurrences were manually counted for each phrase based on Sketch Engine's original results.

As highlighted in table 13, results come mostly from the *Technical expertise + tools* category, with concepts such as computational linguistics and machine learning being present in around a fifth of the documents. Results in this category are a mix of general technological expertise requirements such as programming languages, computer science, and data analysis, and technological requirements that are more language-oriented such as NLP, language technology and language data. Artificial Intelligence is also mentioned in a few texts, with the phrases "conversational ai" and "artificial intelligence" combined appearing in 11 documents. Results in the "Translation and linguistics" category return requirements such as an additional language, language skills and corpus linguistics; while "project management", which is present in around a tenth of the documents, is the only result that did not fit in the two main categories.

Category	Noun+noun / adjective+noun phrase	Number of documents (Occurrences)	% in corpus
Technical expertise	Computational linguistics	23 (27)	20,72%
Technical expertise	Machine learning	22 (25)	19,82%
Technical expertise	Computer science	16 (16)	14,41%
Technical expertise	(Natural) language processing	13 (14)	11,71%
Technical expertise	Programming languages	10 (11)	9,01%
Technical expertise	Data analysis	9 (10)	8,11%
Technical expertise	Language technology	8 (9)	7,21%
Technical expertise	Analytical skills	8 (8)	7,21%
Technical expertise	Programming skills	7 (7)	6,31%
Technical expertise	Version control (systems/git)	6 (7)	5,41%
Technical expertise	Regular expressions	6 (6)	5,41%
Technical expertise	Language data	6 (8)	5,41%
Technical expertise	Conversational AI	5 (5)	4,50%
Technical expertise	Artificial intelligence	4 (6)	3,60%
Translation and linguistics	Additional language	7 (7)	6,31%
Translation and linguistics	Language skills	6 (6)	5,41%
Translation and linguistics	Corpus linguistics	4 (6)	3,60%
Miscellaneous	Project management	13 (16)	11,71%

Table 13: Noun + Noun and Adjective + Noun phrases in the "Required qualifications" section of the WEBCTRL corpus categorized

Although some of the concepts teeter the line between the two categories, such as language technology and language data, the addition of these phrases to the *Translation and linguistics* category would not be enough to balance the scales. Moreover, it could be argued that finding concepts that are hard to place in a technological or language-oriented category is an endeavor closely tied with the analysis of the job profile at hand.

4.2.3 Job functions

The first step in the analysis of the “Job functions” section of the corpus was categorization, as with the “Required qualifications” section. Here, the categories that were previously defined in the UPSKILLS Project, namely *Linguistics, research- and technology-focused tasks* (LRT) and *General tasks*, were deemed appropriate for WEBCTRL results too. Table 14 contains the most common collocates found in the analysis paired with context and collocate analysis of most of the occurrences of each lemma. Lemmas highlighted in yellow belong to the *General tasks* category, while those highlighted in blue belong to the *Linguistics, research and technology* category.

As shown in table 14, most of the results highlighted in blue imply a relatively high degree of technological involvement. Data and quality are the two most common collocates in this category. They encompass tasks such as data analysis for natural language processing and quality control of the output of specific tools. There are also roles which require carrying out research or development of language models, and in some cases the knowledge in understanding and utilizing these tools is also required, such as in the sentences where “AI” is present. Instead, the less technology-focused tasks in these categories have to do with translation. Among them we find ‘providing support as a project manager’ or ‘providing translation first-hand’[13], dealing with localization vendors or taking localization decisions (presumably as a project/vendor manager).

The *General tasks* category highlights teamwork as a very important component of the job, with lemmas such as team and work referring to collaboration, working in cross-functional teams or in dynamic environments. It is interesting to note that results such as product and team paint a complete picture of project development together with lemmas from the previous category. From research and development of the product to improving and managing the product, to meetings with sales team to market the product. In general this category may highlight managerial tasks fit for project, vendor or product management.

[13] These two sentences may not be the exact ones found in the corpus

Category	Lemma	Number of documents (Occurrences)	% in corpus	Collocate analysis / Context
LRT	data	55 (124)	50%	*Annotate quality* data; Perform data *analysis*; improve data *output*
LRT	quality	36 (69)	32%	Perform quality checks/control, guarantee quality of (tools; training; output)
LRT	language	35 (60)	32%	Annotate or tag *natural language; work on natural language* processing; Implement and fine tune language *models*
LRT	process	34 (57)	31%	*Improve* processes and *tools;*
LRT	development	29 (35)	26%	Of *AI*-related products; Of *ontologies*
LRT	analysis	28 (36)	25%	Handle *data* or *language* analysis *requests*
LRT	Content	26 (61)	23%	tech for content *creation;* work with sensitive content (*adult, religious...)*
LRT	model	25 (48)	23%	*Train* and *test* the *performance* of *generative* models
LRT	tool	25 (26)	23%	*Internal* and various types of
LRT	ai	24 (62)	22%	Work with *generative* or *conversational* ai *models*
LRT	research	23 (39)	21%	*Conduct* (*internal*) research to *develop* training materials etc.
LRT	training	21 (32)	19%	*Develop* training *models*; Provide training and *feedback* to coworkers
LRT	performance	19 (28)	17%	*Monitor* or *analyze* system or performance *improvements;*
LRT	technology	19 (20)	17%	*NLP,* various other technologies
LRT	application	17 (26)	15%	*development* and *implementation* of A.
LRT	annotation	16 (35)	14%	Perform *data* annotation according to the *guidelines*
LRT	translation	14 (24)	13%	oversee translation *projects*
LRT	machine	11 (18)	10%	Related to machine *learning*
LRT	localization	6 (21)	5%	deal with localization *vendors,* offer l. *support;* take l. *decisions*
General tasks	team	66 (129)	59%	*Collaborate* (with)*, lead* or *support cross-functional, internal* team *members*
General tasks	project	46 (109)	41%	*Ensure* and *coordinate* project *assignment* and *delivery*
General tasks	work	34 (43)	31%	work collaboratively in a dynamic environment
General tasks	product	30 (41)	27%	Product *managers* or "improve" the product
General tasks	solution	29 (42)	26%	Implement *scalable* solutions
General tasks	issue	27 (44)	24%	*Identify* and *address potential* or *sensitive* issues
General tasks	customer	26 (46)	23%	*Monitor* and *analyze* customer *requests* and *feedback*
General tasks	improvement	26 (39)	23%	(Same as process)
General tasks	client	16 (34)	14%	attend client *meetings* with sales *team*; lead client *engagements*

Table 14: Analysis of the top 28 lemmas in the "Job functions" section of the WEBCTRL corpus

4.2.4 Job title

As mentioned in section 3.2.3 and in the UPSKILLS project, job titles are not analyzed following the same method of analysis as the other sections, but instead adopting a more impressionistic/visual method. Figure 4 contains a word cloud based on the "job title" section of the corpus. Table 15 contains a frequency list of the most common 2-3 grams present in the aforementioned section. Results are categorized based on the most relevant keyword of the job title in question.

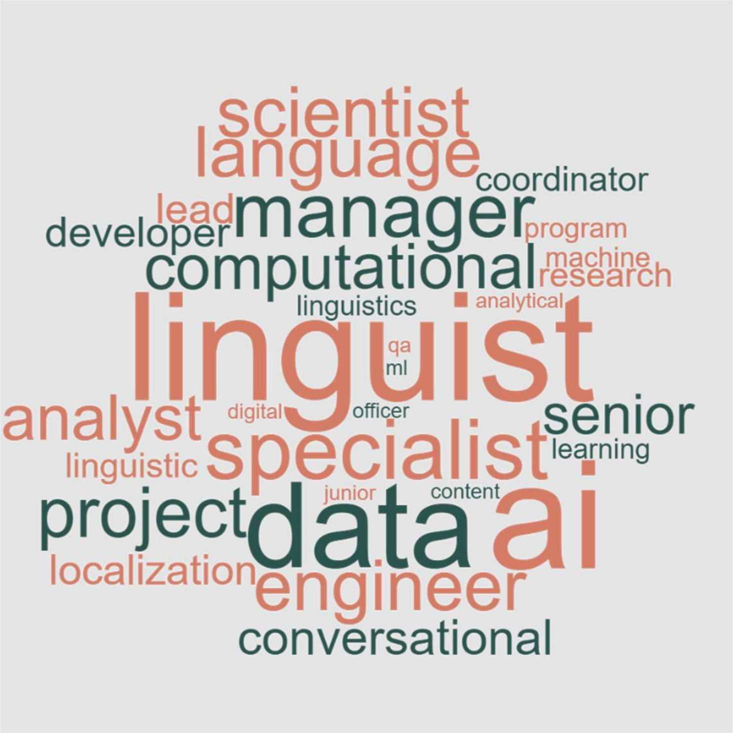


Figure 4: Word cloud based on the "Job title" section of the WEBCTRL corpus

Keyword	Job Title	Frequency	% in corpus
Linguist	Linguist	10	9,01%
	Computational Linguist	7	6,31%
	Analytical Linguist	3	2,70%
AI	Conversational AI [Analyst/Lead/Designer etc.]	7	6,31%
	AI developer	3	2,70%
Manager	Project manager	6	5,41%
	Program manager	3	2,70%
Other	Data scientist	5	4,50%
	Language specialist	5	4,50%
	Machine learning engineer	3	2,70%

Table 15: Categorized most frequent job titles in WEBCTRL

The *Linguist* keyword is the main keyword in both the word cloud and the frequency list. After manually inspecting the job ads for the linguist position, the results were more or less split between a “traditional” linguist providing their expertise and knowledge in various technological settings (data analysis, business development); and a more “technological” linguist involved in prompt engineering, natural language processing and other related matters. In both cases, a certain degree of technological knowledge and involvement is required, and no instances were detected where a linguist could completely do without IT skills.

AI is also heavily featured in the data. Most of the positions specifically have to do with conversational AI, and most titles are inherently tech-based such as (team) lead or analyst. Managerial roles are also represented in the form of project manager and community manager.

It is interesting to note that the *Data* keyword seems to be of high relevance in the word cloud, however it is only featured in the “data scientist” role in the frequency list. This could indicate that data-related tasks are present in most job titles, even those who do not expressly mention data in their job title or that are not typically associated with them. This could also point towards the data scientist position being very central since this was not explicitly targeted during the corpus construction phase.

The *Other* category is composed of mainly tech-centric jobs such as “Machine learning engineer” and “data scientist”. “Language specialist” is also present in this category, which could represent a role where language competencies are of the utmost importance, while technical skills could represent a “nice-to-have”.

The next sub-chapter will offer an overview of results from the analysis of the UPSKILLS corpus, which was carried out from scratch for this thesis, and will compare results between the WEBCTRL corpus and the UPSKILLS corpus to find differences, similarities, and emerging trends.

4.3. Results – UPSKILLS vs. WEBCTRL

This section presents results from both the WEBCTRL and the UPSKILLS corpora, and compares the results of the two while focusing on emerging trends and key differences. It was decided not to dedicate an entire subchapter to UPSKILLS’ corpus analysis because a thorough analysis is already present in Ferraresi et.al (2021) and is freely available online. Moreover, although analysis was carried out from scratch, the differences with the analysis of the UPSKILLS projects are minimal. The tables include thorough results for both corpora, and key results and differences are discussed as the analysis progresses.

The results will be presented in the same order as in the previous sub-chapters, and tables will contain results from the WEBCTRL corpus (with the header highlighted in orange, on the left) and the UPSKILLS corpus (with the header highlighted in blue, on the right). Between the results of the two corpora will be a percentage difference column (of WEBCTRL compared to UPSKILLS). Results are sorted like in the previous sections of the thesis, with matching results from UPSKILLS being placed in the same row of the specific WEBCTRL result.

4.3.1 Required qualifications – Formal education requirements

Table 16 illustrates that the order of the most requested degrees remains consistent across WEBCTRL and UPSKILLS. A bachelor's degree is still the most requested academic requirement, followed by a master's degree and finally by a PhD. However, there is a sharp decline in the necessity for both a bachelor's degree and a master's degree. In the UPSKILLS corpus, a bachelor's degree was explicitly mentioned in half of the documents (50%), whereas in the WEBCTRL corpus this requirement is present in only 37% of the documents, a difference of 13%. The same percentage difference is observed for a master's degree. Finally, PhDs are mentioned in 10% of the documents in the WEBCTRL corpus, a difference of 3% with UPSKILLS. This could point towards there being a need for both junior positions with less degree requirements and highly specialized ones, for which previous academic research is needed.

Degree level	Number of documents	% in corpus	% DIFF	Degree level	Number of documents	% in corpus
Bachelor's degree	41	37%	**-13%**	Bachelor's degree	99	50%
Master's degree	20	18%	**-13%**	Master's degree	62	31%
PhD	11	10%	**3%**	PhD	13	7%
TOTAL	72 out of 111	64%	**-24%**		174 out of 197	88%

Table 16:Degree frequency in WEBCTRL corpus and UPSKILLS corpus

Shifting the focus on to degree types and subjects, table 16 shows how there are no substantial differences among the two corpora. Most of the percentage differences are less than one percentage point, and the maximum difference between the frequency of two degree types is less than 4%. Linguistics remains the most requested degree, while computer science is slightly more requested in the WEBCTRL corpus compared to UPSKILLS. Computational linguistics has declined a bit in comparison to the UPSKILLS corpus, while Mathematics and literature are two degrees that were not present in the UPSKILLS corpus. Perhaps part of the need for a degree in computational linguistics has shifted to even more STEM oriented fields such as mathematics.

Degree in	Number of documents	% in corpus	%DIFF	Degree in	Number of documents	% in corpus
Computer science	10	8,93%	**1,31%**	Computer science	15	7,61%
Translation	3	2,68%	**0,65%**	Translation	4	2,03%
Information science / Information systems	3	2,68%	**0,14%**	Information science / Information systems	5	2,54%
Speech Science / Information Science / Applied Science	4	3,57%	**-0,49%**	cognitive science, library science, data science, social science	8	4,06%
Language Technologies	2	1,79%	**-0,75%**	Language Technologies	5	2,54%
Linguistics	24	21,43%	**-0,91%**	Linguistics	44	22,34%
Computational Linguistics	9	8,04%	**-3,64%**	Computational linguistics	23	11,68%
Mathematics	4	3,57%				
Literature	3	2,68%				

Table 17: Comparison of the degree types in WEBCTRL corpus and UPSKILLS corpus

4.3.2 Required qualifications – Experience and Knowledge

Starting by analyzing the years of experience required by companies in WEBCTRL and UPSKILLS, we can see how jobs requiring 1-3 years of experience are slightly more common in the WEBCTRL corpus, while jobs requiring 4-5 years of experience have suffered a slight decline. It can also be noted that in WEBCTRL, no job advertisement requires more than 5 years of experience, while four ads in UPSKILLS require 8 years of experience. While it must be remembered that WEBCTRL limited the selection of job advertisements asking for several years of experience, due to the nature of the job profile at hand (recent graduates with translation and technology skills) and considering that UPSKILLS did not apply any such restrictions; the data in table 18 could point towards there being more entry level positions in comparison to when the UPSKILLS project was carried out.

Years of Experience	Number of Documents (Occurrences)	% in corpus	% DIFF	Years of Experience	Number of Documents (Occurrences)	% in corpus
1	6 (7)	5,41%	**2,36%**	1	6 (6)	3,05%
2	16 (18)	14,41%	**3,25%**	2	22 (23)	11,17%
3	12 (12)	10,81%	**3,70%**	3	14 (16)	7,11%
4	3 (3)	2,70%	**-0,85%**	4	7 (13)	3,55%
5	6 (6)	5,41%	**-2,21%**	5	15 (17)	7,61%
				8	4 (4)	2,03%

Table 18: Comparison of the years of experience required in WEBCTRL and UPSKILLS

Due to the high amount of results stemming from the analysis of the lemma “experience”, the following results table will be broken down into two tables. The first will only contain collocates which are present in both corpora sorted by percentage difference and then by percentage in corpus and will highlight what has changed in concepts present in WEBCTRL and UPSKILLS. All the collocates that are not present in both corpora will be put in another table, sorted first by field and then by percentage in corpus, to focus on new requirements which may have emerged since the UPSKILLS project or those which are instead declining compared to the past.

In table 19 we can see that the collocate “working with” is much more present in the WEBCTRL corpus, with an increase of almost 20% compared to UPSKILLS. Although it could be argued that this collocate is not very telling in itself, it may indicate a heightened need for practical know-how or knowledge of specific software and operations. It is also interesting to note that the contextual information of the collocate *working with* slightly differs across corpora. While UPSKILLS is mainly focused on the knowledge of languages or language data, WEBCTRL is more focused on diverse types of data such as data for machine learning, and also includes the ability to work with vendors.

Field	Collocate	Number of documents (Occurrences)	% in corpus	Additional information	% DIFF	Collocate	Number of documents (Occurrences)	% in corpus	Additional Information
Other	Working with/in/on	40 (51)	36,04%	data (ML/language/text/ speech), vendors	17,26 %	Working with	37 (45)	18,78%	multiple languages / language data
Technical expertise	Python	13 (15)	11,71%		7,14%	Python	9 (9)	4,57%	
Technical expertise	Programming	14 (16)	12,61%	as in programming languages	6,52%	Programming	12 (14)	6,09%	Python, a p. language // languages
Technical expertise	Machine	11 (11)	9,91%	Learning	5,85%	Machine Learning	8 (10)	4,06%	
Other	Management	16 (25)	14,41%	project/product/ account/product	5,27%	Managing	18 (24)	9,14%	teams, projects
Technical expertise	Data	25 (36)	22,52%	analyze/annotate/ types of data	5,26%	Data	34 (56)	17,26%	language data
Translation and linguistics	Translation	11 (17)	9,91%		3,31%	Translation	13 (16)	6,60%	
Translation and linguistics	Localization	11 (21)	9,91%		2,80%	Localization	14 (21)	7,11%	
Technical expertise	Linux	7 (8)	6,31%		1,74%	Linux	9 (9)	4,57%	
Translation and linguistics	Ontologies	4 (9)	3,60%		1,57%	Ontologies	4 (8)	2,03%	
Technical expertise	Software	10 (15)	9,01%	developing / types of	0,89%	Software	16 (24)	8,12%	pieces of software (e.g. Cogito studio) // e.g. transcription systems
Technical expertise	Systems	7 (11)	6,31%	AI/types of	-0,80%	Systems	14 (15)	7,11%	pieces of software (e.g. Cogito studio) // e.g. transcription systems
Translation and linguistics	Linguistics	19 (27)	17,12%		-2,17%	Linguistics	38 (45)	19,29%	
Technical expertise	Tools	8 (10)	7,21%	localization/types of	-3,45%	Tools	21 (24)	10,66%	command line tools, marketing automation tools
Other	Research	5 (9)	4,50%	processes/market/ industry	-5,14%	Research	19 (26)	9,64%	research processes / research e. / e. worth research
Technical expertise	Computational	7 (9)	6,31%	Linguistics	-5,37%	Computational	23 (26)	11,68%	linguistics // NLP
Translation and linguistics	Language	18 (34)	16,22%	models/multiple/ foreign/ programming	-9,16%	Language	50 (72)	25,38%	multiple languages / language data / programming languages / foreign languages

Table 19: Comparison of the collocates of "experience" present in the "Required qualifications" section of both WEBCTRL and UPSKILLS

The collocates that have increased their presence in job ads mostly belong to the *Technical expertise* category, which contains terms such as "Python", "Programming", "Data" and "Machine" (referring mostly to machine learning); with all of the mentioned terms boasting an increase of 5% to 7%.

The situation has remained almost unchanged (± 3%) for "Translation" and "Localization", two of the most indicative terms for traditional translation jobs. "Ontology" and "Linux" have also undergone marginal changes. The percentage difference across UPSKILLS and WEBCTRL is not relevant for terms such as "Systems" and "Software" (respectively +0.8% and -0,8%), but there is a noticeable difference in the contextual information. While in the UPSKILLS corpus these terms both referred to CAT Tools or transcription systems, in the WEBCTRL corpus there may have been a shift in meaning, as references to development, AI, and different types of software or systems are now much more common for these two terms.

"Research" is one of the fields to have suffered a sharp decline in popularity in the WEBCTRL corpus, with a percentage difference of more than 5%. The collocate "computational", referring to computational linguistics in both corpora, has also followed a similar trajectory. From the rest of the data, it could be speculated that interest in computational linguistics was replaced by the one in machine learning, seeing their percentages share a very similar relative value. Finally, "language" is the one concept with the highest negative percentage difference, with a difference of almost 10% from the UPSKILLS corpus. This, coupled with previous information, might point towards a heavier shift towards more tech-oriented requirements. This is also suggested by the change in the contextual information, which now includes more references to language models and programming in the WEBCTRL corpus.

As for collocates present in either WEBCTRL or UPSKILLS but not in both corpora, an overview is presented in table 20. Starting from collocates present only in the WEBCTRL corpus, "using" and "environment" are the two top results by percentage in corpus and, although it could be argued that they are a bit general, contextual information does point towards them referring to programming languages or "scholarly" environments. One of the more interesting terms only present in this corpus is "AI", which can be found in more than 11% of the WEBCTRL corpus, while no occurrences were found in the UPSKILLS corpus. This could point towards a need for experience in artificial intelligence only becoming popular in recent years. In general, WEBCTRL-exclusive collocates point towards a renewed interest towards the more technical requirements, although wording could have conditioned some of the results. For example, "project" (as in project management) seems to be exclusive to WEBCTRL, but we have seen that "managing" is also present in UPSKILLS with contextual information related to managing projects and teams.

Field	Collocate	Number of documents (Occurrences)	% in corpus	Additional information
Other	Using	16 (30)	14,41%	various programs/programming languages
Other	Project	12 (16)	10,81%	management/coordination
Other	Communication	10 (11)	9,01%	Skills
Other	Industry	7 (14)	6,31%	
Technical expertise	Environment	15 (17)	13,51%	scholarly/connected/related to skill
Technical expertise	AI	13 (16)	11,71%	generative / AI in general
Technical expertise	Analysis	9 (11)	8,11%	data
Technical expertise	SQL	8 (8)	7,21%	
Technical expertise	Develop	16 (21)	7,21%	Software/Products
Technical expertise	GIT	6 (8)	5,41%	
Technical expertise	Technical	6 (8)	5,41%	skill/fields
Translation and linguistics	Text	9 (11)	8,11%	data/classification

Field	Collocate	Number of documents (Occurrences)	% in corpus	Additional Information
Other	Professional	17 (22)	8,63%	professional e. / e. in professional settings
Other	Writing	14 (20)	7,11%	grammars, SQL, code, documentation
Other	Team	14 (18)	7,11%	managing a team / working with teams
Other	Designing	10 (15)	5,08%	documentation, interfaces
Other	Practical	10 (14)	5,08%	
Technical expertise	Scripting	14 (15)	7,11%	Python, a p. language // languages
Technical expertise	NLP	13 (19)	6,60%	
Technical expertise	Large	8 (9)	4,06%	datasets, quantities of data
Technical expertise	Speech recognition	5 (7)	2,54%	
Translation and linguistics	Annotation	22 (30)	11,17%	
Translation and linguistics	Semantics, syntax, morphology	4 (5)	2,03%	

Table 20:Collocates of "experience" present in the “Required qualifications” section of WEBCTRL only, or UPSKILLS only

The collocate with the highest percentage in corpus present only in the UPSKILLS corpus, and arguably also the most indicative, is “annotation”. It could be speculated that the rise in mentions for AI and the decline in requests for annotation experience may be because employers are starting to annotate data with or for AI, and thus have shifted their needs in this direction. The rest of the collocates exclusive to the UPSKILLS corpus are not very telling, with results such as “writing” including SQL in the contextual information, which is a term that is present in the WEBCTRL corpus, and others like “scripting” or “NLP” bearing a similar meaning to collocates present in both corpora.

Table 21 shows collocates of the lemmas *knowledge* and *understanding* present in both corpora. The analysis does not reveal any glaring differences between WEBCTRL and UPSKILLS, as the maximum increase in percentage in corpus is represented by “data” with an increase of only +2%, and the highest decrease is represented by “semantics”, with a change of -3%.

Field	**Lemmas**	**Collocate**	**Number of Documents (Occurrences)**	**% in corpus**	**Additional Information**	**% DIFF**	**Collocate**	**Number of Documents (Occurrences)**	**% in corpus**	**Additional Information**
Technical expertise	Knowledge / Understanding	data	8 (8)	7,21%	mining, science languages, governance, analysis	**2,13%**	Data	10 (10)	5,08%	data structures, data processing needs
Technical expertise	Knowledge / Understanding	programming language	3 (4)	2,70%		**0,67%**	Programming	4 (5)	2,03%	
Translation and linguistics	Knowledge / Understanding	translation	3 (3)	2,70%	one was translation software	**0,16%**	Localization	5 (5)	2,54%	
Translation and linguistics	Knowledge / Understanding	structural aspects of languages	4 (4)	3,60%	this exact phrasing was found all 4 times	**0,05%**	Structural aspects of language	7 (7)	3,55%	this exact phrasing was found all 7 times
Translation and linguistics	Knowledge / Understanding	linguistics	4 (4)	3,60%		**-0,46%**	Linguistics	8 (8)	4,06%	
Technical expertise	Knowledge / Understanding	Microsoft Office	3 (3)	2,70%		**-1,36%**	Microsoft Office/Tools	8 (8)	4,06%	
Translation and linguistics	Knowledge / Understanding	language	11 (12)	9,91%	mostly programming languages. only 3 occurrences refer to foreign languages or language aspects	**-1,77%**	Language	23 (26)	11,68%	15/23 refer to foreign/additional languages
Translation and linguistics	Knowledge / Understanding	syntax	4 (5)	3,60%		**-1,98%**	Syntax	11 (11)	5,58%	
Technical expertise	Knowledge / Understanding	tools	3 (4)	2,70%	various, technical, seo, statistical	**-2,38%**	Tools	10 (13)	5,08%	CAT tools, software tools
Translation and linguistics	Knowledge / Understanding	semantics	2 (3)	1,80%		**-3,28%**	Semantics	10 (10)	5,08%	

Table 21: Comparison of the collocates of "knowledge" and “understanding” present in the “Required qualifications” section of both WEBCTRL and UPSKILLS

Table 22 shows collocates of *skills* and *abilities* present in both WEBCTRL and UPSKILLS. It is instantly noticeable that the percentage in corpus of most of the collocates in the WEBCTRL corpus has increased compared to UPSKILLS, potentially pointing towards skills and abilities being more of a focal point in job advertisement now than in the past.

Somewhat inverting the trend of the past analyses, the collocates with the biggest increase in percentage in corpus mostly belong to the *Translation and linguistics* category. Although they could be classified as soft skills, it can be argued that communication skills (especially written communication skills) are paramount for a translator or a linguist, even if they can be widely applicable. "Communication" (+12%) and its various types like "verbal" (+8%) and "written" (+7%) all see a significant increase, together with "problem-solving" (+7%) skills from the *Other* field and "Python" (+5%) skills from the *Technical Expertise* field. The other results all display a small increase in percentage in corpus, with contextual information remaining very similar across corpora. The only collocates with a decline in percentage in corpus in the WEBCTRL corpus are "analytical" (-2%), "language" (-2%) and "organizational" (-7%).

Field	Lemmas	Collocate	Number of Documents (Occurrences)	% in corpus	Additional Information	% DIFF	Collocate	Number of Documents (Occurrences)	% in corpus	Additional Information
Translation and linguistics	Skills	Communication	61 (76)	54,95%		12,31%	Communication	84 (97)	42,64%	written and oral/verbal/spoken c. skills
Translation and linguistics	Skills	Written	31 (38)	27,93%	mostly communication	8,64%	written	38 (47)	19,29%	communication
Other	Skills	Problem-solving	13 (14)	11,71%		7,14%	Problem-solving	9 (10)	4,57%	
Translation and linguistics	Skills	Verbal	19 (25)	17,12%	mostly communication	6,97%	verbal	20 (25)	10,15%	communication
Technical expertise	Skills	Python	8 (10)	7,21%		5,18%	Python	4 (5)	2,03%	
Translation and linguistics	Skills	Writing	9 (16)	8,11%	various types of writing (technical, coding, content, business level English etc.)	4,56%	writing	7 (7)	3,55%	presentation / reports
Other	Abilities	Manage	11 (11)	9,91%	multiple priorities	4,33%	Manage	11 (11)	5,58%	mostly overlap with priority/ projects / teams /relationships
Other	Skills	Interpersonal	14 (17)	12,61%		3,47%	Interpersonal	18 (20)	9,14%	
Other	Abilities	Work	25 (28)	22,52%	independently/collaboratively/remotely/under pressure	3,23%	Work	38 (42)	19,29%	independently/under pressure etc.
Other	Skills	Problem (solving)	8 (8)	7,21%		2,64%	Problem (solving)	9 (11)	4,57%	
Other	Skills	(Critical) Thinking	6 (8)	5,41%		2,36%	Thinking	6 (7)	3,05%	critical / negotiation
Technical expertise	Abilities	data	5 (5)	4,50%	use/analyze	1,96%	data	5 (5)	2,54%	interpret / translate
Other	Abilities	multitask	4 (5)	3,60%		1,57%	multitask	4 (4)	2,03%	
Other	Skills	Management	13 (18)	11,71%	Project/time/people management	1,56%	Manage / management	20 (21)	10,15%	manage projects / priorities
Technical expertise	Skills	Research	5 (10)	4,50%		1,45%	research	6 (7)	3,05%	
Technical expertise	Skills	Programming	6 (6)	5,41%		1,35%	programming	8 (9)	4,06%	
Other	Abilities	Learn	5 (6)	4,50%	tools/software	0,95%	Learn	7 (8)	3,55%	new skills/systems/software
Translation and linguistics	Skills	English	18 (21)	16,22%		0,48%	English	31 (36)	15,74%	
Technical expertise	Skills	Computer	6 (6)	5,41%		0,33%	Computer	10 (11)	5,08%	
Other	Abilities	Prioritize	8 (8)	7,21%	tasks	0,10%	Priority / prioritize	14 (14)	7,11%	Manage needs and tasks
Other	Skills	detail	13 (13)	11,71%	As in attention to detail	0,03%	(Attention to) detail	23 (32)	11,68%	
Technical expertise	Skills	Analytical	12 (14)	10,81%		-1,88%	Analytical	25 (33)	12,69%	
Translation and linguistics	Skills	Language	11 (14)	9,91%	mainly foreign	-2,27%	language	24 (31)	12,18%	mostly foreign
Other	Skills	Organizational	12 (12)	10,81%		-6,96%	Organizational	35 (44)	17,77%	

Table 22:Comparison of the collocates of "skills" and “abilities” present in the “Required qualifications” section of both WEBCTRL and UPSKILLS

Finally, tables 23 and 24 show the results of the query targeting noun+noun or adjective+noun phrases. As was done with the analysis of the lemma experience, results will be broken down in two tables. The first only contains results which are present in both corpora and is sorted by percentage difference, while the second shows results only present in one corpus and is sorted by field first and then by percentage in corpus.

Starting out by looking at table 23, we can see that “machine learning” is the phrase with the highest growth in comparison to UPSKILLS (+6%). A small growth is also observed for “additional/second language” (+3%), “version control tools” (+2%), “data analysis” (+2%) and “computer science” (+1%). Among results which have more or less remained unchanged we can find “project management”, “regular expressions” and “programming languages”; while “language data” and “natural language processing” have suffered a very slight decline (both around -1%). “Computational linguistics” is the phrase with the most noticeable drop in frequency (-5%) which, again, might have a relationship with the rise in popularity of machine learning.

Field	Noun+noun / adjective+noun phrase	Number of documents (Occurrences)	% in corpus	% DIFF	Noun+noun / adjective+noun phrase	Number of documents (Occurrences)	% in corpus
Technical expertise	Machine learning	22 (25)	19,82%	**5,61%**	Machine learning	28 (39)	14,21%
Translation and linguistics	Additional language	7 (7)	6,31%	**3,26%**	Second language	6 (8)	3,05%
Technical expertise	Version control (systems/GIT)	6 (7)	5,41%	**1,86%**	Version control systems/tools	7 (7)	3,55%
Technical expertise	Data analysis	9 (10)	8,11%	**1,51%**	Data analysis	13 (15)	6,60%
Technical expertise	Computer science	16 (16)	14,41%	**1,21%**	Computer science	26 (31)	13,20%
Other	Project management	13 (16)	11,71%	**0,54%**	Project management	22 (24)	11,17%
Technical expertise	Regular expressions	6 (6)	5,41%	**0,33%**	Regular expressions	10 (10)	5,08%
Technical expertise	Programming languages	10 (11)	9,01%	**-0,13%**	Programming language	18 (19)	9,14%
Technical expertise	Language data	6 (8)	5,41%	**-1,19%**	Language data	13 (13)	6,60%
Technical expertise	(Natural) language processing	13 (14)	11,71%	**-1,49%**	Natural language processing	26 (27)	13,20%
Technical expertise	Computational linguistics	23 (27)	20,72%	**-4,66%**	Computational linguistics	50 (62)	25,38%

Table 23: Comparison of Noun + Noun and Adjective + Noun phrases in the "Required qualifications" section of WEBCTRL and UPSKILLS corpora

Table 24 offers an overview of phrases exclusive to either WEBCTRL or UPSKILLS. Starting with WEBCTRL, "language technology" and "analytical skills" are the phrases with the highest amount of percentage in corpus (7%), with "programming skills" trailing slightly behind. It is interesting to note that artificial intelligence is mentioned in two phrases, first with "conversational AI" and then with "artificial intelligence". If we combine the percentage in corpus for these two phrases, then AI would have a percentage in corpus of around 8%, more than any other phrase in WEBCTRL. Finally, "corpus linguistics" is the only item belonging in the *Translation and linguistics* field, with a percentage in corpus of around 3%.

As for UPSKILLS, the most common phrase not present in WEBCTRL is "technical concepts" (10% percentage in corpus), which is somewhat misleading, seeing that results from the WEBCTRL corpus mostly belong to the *Technical expertise* field. "Speech recognition", "command line" (as in command line tools), "machine translation" and "software development" are all items from the aforementioned field that are exclusive to UPSKILLS. The results from the *Translation and linguistics* field are mostly general and with plenty of language requirements, aside from "annotation experience" which could switch categories depending on the type of annotation (for AI or language corpora, for example). Finally, "social media" and "customer service" are results from the *Other* field that do not figure in the WEBCTRL corpus.

Field	**Noun+noun / adjective+noun phrase**	**Number of documents (Occurrences)**	**% in corpus**
Technical expertise	Language technology	8 (9)	7,21%
Technical expertise	Analytical skills	8 (8)	7,21%
Technical expertise	Programming skills	7 (7)	6,31%
Technical expertise	Conversational AI	5 (5)	4,50%
Technical expertise	Artificial intelligence	4 (6)	3,60%
Translation and linguistics	Corpus linguistics	4 (6)	3,60%
Field	**Noun+noun / adjective+noun phrase**	**Number of documents (Occurrences)**	**% in corpus**
Technical expertise	Technical concepts	19 (19)	9,64%
Technical expertise	Speech recognition	9 (11)	4,57%
Technical expertise	Command line	8 (8)	4,06%
Technical expertise	Machine translation	6 (7)	3,05%
Technical expertise	Software development	6 (8)	3,05%
Translation and linguistics	Native speaker	18 (28)	9,14%
Translation and linguistics	Target language	8 (10)	4,06%
Translation and linguistics	English language	7 (8)	3,55%
Translation and linguistics	Annotation experience	6 (6)	3,05%
Other	Social media	11 (14)	5,58%
Other	Customer service	6 (6)	3,05%

Table 24: Noun + Noun and Adjective + Noun phrases in the "Required qualifications" section of only WEBCTRL or UPSKILLS

4.3.3 Job functions

For the comparison between the WEBCTRL corpus and the UPSKILLS corpus concerning job functions, the results from the UPSKILLS corpus were categorized using the new categories defined for the WEBCTRL corpus: *LRT* (*Linguistics, research- and technology-focused tasks*) and *general tasks*. As with the analysis of the lemma *experience* and noun + noun or adjective + noun phrases, results will be broken down in two tables. The first table contains results present in both corpora, which are sorted by percentage difference between WEBCTRL and UPSKILLS, while the second shows results only present in one corpus which are sorted by category first and then by percentage in corpus. In both tables, words in italics in the "Contextual information" column represent co-occurring words, while the content of the column itself tries to provide information in a colloquial way while trying to fit in co-occurring words to provide additional information.

As can be seen in table 25, "model" and "performance" are the two lemmas with the highest increase in percentage in corpus, with 8% and 6% respectively. Both lemmas seem to mostly refer to different tasks related to generative models. Just below, we can find "data", "annotation" and "team", which share both a similar percentage increase (4% for data, 3% for the other two) and a similar meaning of the lemmas across corpora. It is interesting to note that while annotation experience follows a downward trend, employers still need employees to carry out tasks related to annotation. It can be argued that this is in relation to annotation for AI or language models, and therefore employers seek candidates with a background in these sectors rather than specifically in annotation. Another reason could be that annotation tasks are becoming more common or basic, and previous experience can be foregone.

"Content" is an interesting lemma, as it has undergone a 3% increase and a drastic shift in meaning. While content in the UPSKILLS corpus seems to refer to content annotation and (the creation of) learning content, the same lemma in WEBCTRL refers mostly to content creation (to release on YouTube, TikTok or similar platforms) and accepting working with sensitive content (of religious and adult nature).

As for lemmas which have mostly retained their percentage in corpus across corpora, we can find "research", "development", "customer" and "translation", which all fall across a range of ±1%. The lemma "development" has slightly shifted to refer to the development of AI products and ontologies, while in UPSKILLS it mostly referred to research and development in general, and to the development of linguistic databases. The other three lemmas have mostly retained their meaning.

Venturing into lemmas which have suffered a slight decrease in their percentage in corpus, we find that “analysis” (-2%), “localization” (-3%), “project” (-4%) and “tools” (-5%) have all mostly retained their meaning with a few exceptions. It is possible to notice that “localization” refers to more tasks in the WEBCTRL corpus, such as dealing with localization vendors or taking localization decisions. The opposite is true of the lemma “tool”: in the UPSKILLS corpus it refers to software, technological, NLP and internal tools, while only the latter is present in WEBCTRL.

The lemmas with the sharpest drop in percentage in corpus are “quality”, “language”, and “client”. “Quality” has suffered a 7% decrease, and the contextual information is very similar across corpora. The lemma “language” has undergone an even higher 12% decrease in percentage in corpus, and its meaning in WEBCTRL is completely tech-oriented, whereas in the UPSKILLS corpus it was mentioned that language referred also to jobs such as language manager or specialist. Finally, the lemma “client” has suffered a whopping 25% decrease, probably pointing towards a sharp decline in client-facing tasks for this job profile.

Category	Lemma	Number of documents (Occurrences)	% in corpus	Contextual information	% DIFF	Lemma	Number of documents (Occurrences)	% in corpus	Contextual information
LRT	model	25 (48)	23%	*Train* and *test* the *performance* of *generative* models	**8%**	Model	30 (54)	15%	build language models; train models
LRT	performance	19 (28)	17%	*Monitor* or *analyze* system or performance *improvements;*	**6%**	Performance	22 (39)	11%	analyze, test or improve (system/product) performance
LRT	data	55 (124)	50%	*Annotate quality* data; Perform data *analysis*; improve data *output*	**4%**	Data	91 (202)	46%	analyze data; collect data; participate in data *collection* and *annotation projects*
LRT	annotation	16 (35)	14%	Perform *data* annotation according to the *guidelines*	**3%**	Annotation	23 (41)	11%	perform data annotation; update/create annotation guidelines
LRT	Content	26 (61)	23%	tech for content *creation;* work with sensitive content (*adult, religious...)*	**3%**	Content	39 (75)	20%	*annotate C.;* create learning content
General tasks	team	66 (129)	59%	*Collaborate* (with)*, lead* or *support cross-functional, internal* team *members*	**3%**	Team	111 (249)	56%	work with or support teams (e.g. product team, engineering team, project team, development team); collaborate with team members
LRT	research	23 (39)	21%	*Conduct* (*internal*) research to *develop* training materials etc.	**1%**	Research	41 (66)	20%	conduct research; support or participate in research and development
LRT	development	29 (35)	26%	Of *AI*-related products; Of *ontologies*	**0%**	development	51 (66)	26%	*research* and D. ; D. of linguistic *databases* with *engineers*
General tasks	customer	26 (46)	23%	*Monitor* and *analyze* customer *requests* and *feedback*	**-1%**	Customer	47 (78)	24%	provide customer *support* for a range of *internal* and *data* c.
LRT	translation	14 (24)	13%	oversee translation *projects*	**-1%**	Translation	28 (44)	14%	*machine* translation; t. *projects*
LRT	analysis	28 (36)	25%	Handle *data* or *language* analysis *requests*	**-2%**	analysis	54 (65)	27%	perform *error/data A.* in the *linguistic/translation* field
LRT	localization	6 (21)	5%	deal with localization *vendors,* offer l. *support;* take l. *decisions*	**-3%**	Localisation	16 (28)	8%	ensure consistency of localizations *projects*
General tasks	project	46 (109)	41%	*Ensure* and *coordinate* project *assignment* and *delivery*	**-4%**	Project	89 (170)	45%	manage, lead or oversee projects
LRT	tool	25 (26)	23%	*Internal* and various types of	**-5%**	Tools	56 (64)	28%	improve or develop software/ technological/NLP tools; use internal tools
LRT	quality	36 (69)	32%	Perform quality checks/control, guarantee quality of (tools; training; output)	**-7%**	Quality	77 (101)	39%	perform quality controls/assurance; improve quality (of tools/data output)
LRT	language	35 (60)	32%	Annotate or tag *natural language; work on natural language* processing; Implement and fine tune language *models*	**-12%**	language	87 (138)	44%	tasks in *natural l. processing/understanding*; be a l. *manager/specialist*
General tasks	client	16 (34)	14%	attend client *meetings* with sales *team*; lead client *engagements*	**-25%**	Clients	78 (112)	39%	interact with clients; provide customer service or support; participate in meetings with clients; assist clients in developing their business; make sure that customer experience is smooth; manage client accounts

Table 25: Comparison of the top lemmas in the "Job functions" section of the WEBCTRL and UPSKILLS corpora

Looking at table 26, we can see that one of the most indicative results present exclusively in the WEBCTRL corpus is “AI”, with a percentage in corpus of 22%, highlighting the rise of artificial intelligence in both requirements and job functions. Continuing with lemmas present only in WEBCTRL, “training”, “technology”, “application”, and “machine” all hover around 15% to 20% and belong to the *LRT* category. “Training” is categorized as *LRT* because it refers to the development or training of models, however, it can also refer to providing training and feedback to coworkers which is more of a general task. The lemma “machine” refers to machine learning, continuing the upward trend of this concept. Only five items are present in the *general tasks* category exclusive to the WEBCTRL corpus. They are not too indicative, and all five hover around 20 to 30 percentage in corpus.

As for lemmas present only in the UPSKILLS corpus, “information” and “transcription” are the only two items belonging to the *LRT* category. “Information” (17%) mainly refers to information retrieval, which does not seem to be a task candidates will need to carry out anymore, since it is not present in the WEBCTRL corpus. “Transcription” is also absent from WEBCTRL but was present in only 7% of UPSKILLS’ documents to start. As for *general tasks*, the three lemmas belonging in this category are “report”, “vendors”, and “materials”. The lemma “report” (18%) mainly refers to providing project reports and does not seem to be referenced in the WEBCTRL corpus. Although the lemma “vendors” does not appear in WEBCTRL, it is a co-occurring word of the lemma “localization”, referring to dealing with localization vendors.

Category	Lemma	Number of documents (Occurrences)	% in corpus	Contextual information
LRT	process	34 (57)	31%	*Improve* processes and *tools;*
LRT	AI	24 (62)	22%	Work with *generative* or *conversational* ai *models*
LRT	training	21 (32)	19%	*Develop* training *models*; Provide training and *feedback* to coworkers
LRT	technology	19 (20)	17%	*NLP,* various other technologies
LRT	application	17 (26)	15%	*development* and *implementation* of A.
LRT	machine	11 (18)	10%	Related to machine *learning*
General tasks	work	34 (43)	31%	work collaboratively in a dynamic environment
General tasks	product	30 (41)	27%	Product *managers* or "improve" the product
General tasks	solution	29 (42)	26%	Implement *scalable* solutions
General tasks	issue	27 (44)	24%	*Identify* and *address potential* or *sensitive* issues
General tasks	improvement	26 (39)	23%	(Same as process)

Category	Lemma	Number of documents (Occurrences)	% in corpus	Contextual information
LRT	Information	52 (69)	17%	extract information; participate in information *retrieval*
LRT	Transcription	15 (22)	7%	perform phonetic transcription; provide
General tasks	Report	37 (55)	18%	provide written reports; write up project reports
General tasks	Vendors	21 (33)	10%	work with or support external vendors; assess vendors’ performance
General tasks	Materials	28 (54)	7%	*categorize* or *scan* materials

Table 26:Analysis of the top lemmas in the "Job functions" section present only in WEBCTRL or UPSKILLS

4.3.4 Job Title

For the comparison of job titles between the WEBCTRL corpus and the UPSKILLS corpus, a table will first illustrate the differences between the most common 2-3 grams between the two corpora. Since the frequency numbers are very small, percentage in corpus will not be calculated, and the percentage difference column will be replaced by an occurrences difference column. While WEBCTRL and UPSKILLS differ in size (around 75k tokens vs. 107k tokens), it was decided that it would be more understandable to directly reference frequencies. The difference in tokens between the two corpora will be kept in mind during the analysis of the results.

Table 27 contains the most common job titles divided by keyword. The keyword acts as a sort of "container" for related job titles, and was given to each title after analysis. The table is ordered by frequency in the WEBCTRL corpus, while the job title in the UPSKILLS section matches the WEBCTRL one. Job titles that are not present in WEBCTRL are listed at the end of the table.

Keyword	Job Title	Frequency	Freq. Diff	Keyword	Job Title	Frequency
Linguist	Linguist	10	2	Linguist	Linguist	8
	Computational Linguist	7	-9		Computational Linguist	16
	Analytical Linguist	3	-3		Analytical Linguist	6
AI	Conversational AI [Analyst/Lead/Designer etc.]	7				
	AI developer	3				
Manager	Project manager	6	-5	Manager	Project Manager	11
	Program manager	3				
Other	Data scientist	5	0	Data	Data Scientist	5
	Language specialist	5				
	Machine learning engineer	3				
					Associate Linguist	11
					Language Manager	3
					Localization Project Manager	3
				Other	Speech Scientist	4
					Language Analyst	3
					Project Coordinator	3
					Data Linguist	4
					Data Analyst	3

Table 27:Comparison of most frequent job titles between WEBCTRL and UPSKILLS

Table 27 shows various degrees of decline for job titles present in both corpora. The linguist job title more or less has the same frequency with only two more occurrences, while computational linguist has 9 fewer entries in the WEBCTRL corpus. As seen throughout this comparison, computational linguistics seems to be in less demand for either machine learning or more STEM oriented fields, titles and skills.

The *AI* keyword is a new addition and is present only in the WEBCTRL corpus, highlighting just how fast the need for AI-related skills is rising. Counting all instances of job titles involving conversational AI and the AI developer position we have 10 occurrences in this keyword, just two less than the *Data* keyword in the UPSKILLS corpus which has disappeared from WEBCTRL. The only job title still present in WEBCTRL is data scientist with five occurrences in both corpora and, considering the difference in tokens, could be said to be slightly on the rise. Since this was the only data-related job title, the *Data* keyword was removed. Entries for the *Manager* keyword are also in decline or not present in WEBCTRL, while the job titles in the *Other* keyword do not match across corpora with the exception of data scientist.

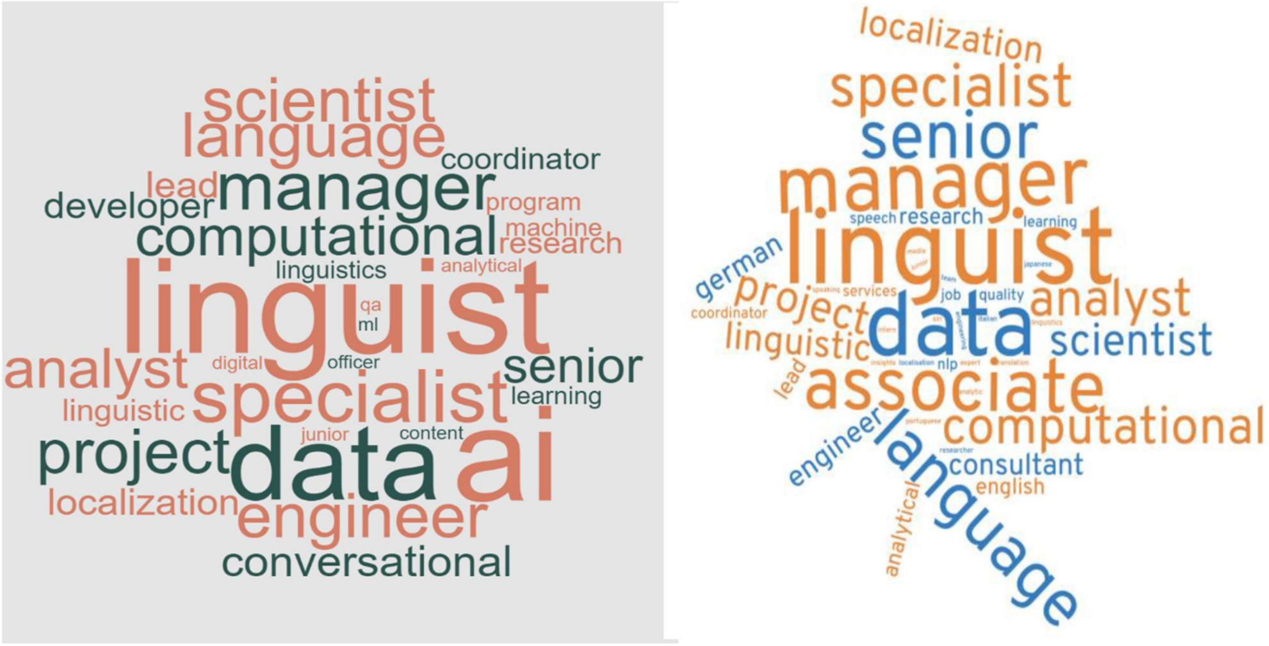


Figure 5: Comparison of Word cloud based on the "Job title" section of WEBCTRL (on the left) and UPSKILLS (on the right)

Figure 5 contains both the word cloud of WEBCTRL (on the left) and the word cloud of UPSKILLS, on the right. It is noticeable that the main keywords have mostly remained the same. Linguist is still the main focus of the word cloud, although it may also be by virtue of being a keyword during WEBCTRL's corpus construction phase. Words like data, manager, computational and language retain approximately the same importance. One important addition is AI, which seems to be the second biggest word in the WEBCTRL word cloud, and confirms its upward trend already seen in this chapter. In general, the WEBCTRL word cloud seems to be even more tech-oriented than its UPSKILLS counterpart, with words such as developer and engineer having a certain prominence. One final interesting aspect is how the keyword senior has become smaller in the WEBCTRL word cloud, which also contains the word junior, albeit smaller. This could point towards jobs in this sector becoming more accessible to recent graduates.

5. Conclusion and Future Studies

This thesis was concerned with the creation and analysis of a corpus of job advertisements for skill identification. First, the WEBCTRL corpus was created with the same goal of the UPSKILLS report, that is to examine and analyze advertisements for jobs that require a combination of linguistic and technological skills and gauge the type of requirements and tendencies of the market related to this job profile. Secondly, the structure and analysis of WEBCTRL were designed to facilitate a comparison with the UPSKILLS corpus and see what exactly has changed since its creation.

Chapter 1 introduced the UPSKILLS corpus and the scope of this work, while chapter 2 introduced concepts necessary to fully grasp the context of the thesis while reviewing previous literature germane to the topic at hand. Chapter 3 dealt with the methodology used to build and annotate WEBCTRL, using a similar method to that of the UPSKILLS corpus but expanding upon its scope and annotation methods. Moreover, it described the WEBCTRL corpus and touched upon the analysis methodology which closely resembles UPSKILLS in order to maintain comparability, although it sports a few differences. Finally, Chapter 4 delved into the result of the WEBCTRL corpus analysis and of comparison between the two corpora.

The findings of this thesis seem to corroborate those of Ferraresi et.al (2021) on a macroscopic level. In the UPSKILLS report four main categories of skills and competences were deemed particularly salient: data and research skills, technical skills, language and linguistics disciplinary knowledge, and communication, interpersonal and organizational skills. All of these requirements retain their importance in WEBCTRL and remain crucial to the job profile tackled in this work. The most apparent trend is that of tech-related requirements still being highly requested by employers, while linguistics-related requirements are suffering a moderate downfall. For example, artificial intelligence was completely absent in UPSKILLS, while it is prominently featured in WEBCTRL.

From the analysis carried out in this thesis, it is possible to draw an overview of the "ideal" job profile in this sector and of the duties they will carry out. A candidate should have around two years of experience and be in possession of a linguistics degree and/or a more tech-oriented degree in a field such as computational linguistics or computer science. They would be experienced with data, linguistics, programming, and management and be skilled in communicating, problem-solving and analysis with a keen eye for detail. The disciplinary concepts they excel in would include computational linguistics, machine learning, NLP and project management. Finally, their day-to-day job would be comprised of annotating and analyzing data, coordinating projects, quality control, language model fine-tuning and product development.

Degree in	Experience in	Skilled in	Disciplinary concepts	Job functions
• Linguistics • Computer science • Computational linguistics • Mathematics • Speech science / information science / applied science	• Data • Linguistics • Language • Management • Programming	• English • Communicating • Problem-solving • Analysis • Attention to detail	• Computational linguistics • Machine learning • Computer science • Natural language processing • Project management	• Annotate/analyze data • Coordinate projects • Quality control of tools or training output • Work with NLP or fine-tune language models • Develop products or ontologies

Degree in	Experience in	Skilled in	Disciplinary concepts	Job functions
• Linguistics • Computational linguistics • Computer science • Cognitive/data […] science • Language technologies	• Language • Linguistics • Data • Computational linguistics • Annotation	• Communicating • Management • (Foreign) Languages • English • Analysis	• Computational linguistics • Machine learning • Natural language processing • Computer science • Project management	• Collect / Analyze data • Be a langauge manager/specialist • Manage or oversee projects • Perform quality assurance • Customer support

Table 28: Top 5 results for each category in WEBCTRL (orange) and UPSKILLS (blue) in decreasing frequency order

The above job profile was made by considering the top 5 results in each of the main categories by percentage in corpus and excluding general results such as "working in a team", as they were not deemed characteristic of this specific role. By applying the same concept to the UPSKILLS corpus and comparing it with WEBCTRL, we can assume that the job profile has not drastically changed since the UPSKILLS corpus was constructed, and that the profile seen in the UPSKILLS report remains valid. It must be said though, that the job profile seems to be shifting towards technology. The experience that was once required for annotation tasks has now changed to a requirement in programming, experience with data seems indispensable now, and the job duties the candidates are expected to carry out have begun to shift away from language-oriented tasks to tasks.

Of course this analysis does not capture all of the details present in the results section, and is meant as a device to concisely respond to the first question of the thesis while also presenting the main differences to the UPSKILLS corpus. It is possible to speculate from the data that the linguist position is still in demand, but it is increasingly shifting towards integration with research skills and, most importantly, technological skills. It is paramount for linguists to adapt to the necessities of the market and broaden their skillset beyond their traditional translation and linguistic background.

The contextual information of the collocates is another important aspect in the context of comparing the two corpora. In the WEBCTRL corpus, contextual information lends a technological emphasis to keywords and collocates that, in the UPSKILLS corpus, carried a stronger focus on language. For example, the "data" collocate present in the Required Qualifications section mostly refers to language data in the UPSKILLS corpus, while WEBCTRL puts more emphasis on analysis and annotation of data. Language is a collocate that was found in multiple sections of the two corpora, understandably so given the field of analysis, and is another example of the above phenomenon. In the Required Qualifications section, while analyzing collocates of "knowledge" and "understanding", this collocate mostly refers to foreign languages in the UPSKILLS corpus (15 out of 23 documents), while in the WEBCTRL corpus 9 out of 12 documents refer to programming languages instead.

It may be possible for future studies to hold this work as a starting point for a more in-depth analysis, while keeping in mind that there are still improvements to be made. Starting from corpus construction, it may be beneficial to forego the semi-automatic corpus construction methods of this paper and employ more sophisticated web scrapers. This would lead to the creation of a larger corpus, although it remains to be seen if such corpus would remain as topical as UPSKILLS or WEBCTRL; as some text selection criteria may be too arbitrary and pose a challenge to fully automatic corpus building methods. As for the next step, annotating the corpus, future work could take advantage of the foundations laid in this thesis in separating mandatory qualifications from preferred qualifications and conduct a more thorough investigation on this aspect. While the tools and methods used to annotate WEBCTRL are slightly more streamlined than those employed for the UPSKILLS corpus, they remain a big improvement opportunity. Further work would greatly benefit from more advanced annotation tools that can improve annotation speed and facilitate more thorough annotation. Finally, it may be relevant for future studies to continue investigating jobs that require both translation and technological skills to verify that the trends highlighted by this analysis (an upwards trend for technological requirements, and a downward trend for linguistic requirements) will continue to move in the same direction.

Annex 1 – Original UPSKILLS sources for corpus construction

wonderflow.bamboohr.com
linkedin.com
indeed.com/viewjob
toplanguagejobs.com/jobs/
seekorswim.com
reed.co.uk/jobs
boards.greenhouse.io
jobs.careers.microsoft.com
talent.com
careers-page.com/*/job
monster.com
careers.mitre.org
careers-maslansky.icims.com
linguistlist.org
jora.com
summalinguae.bamboohr.com
eu-careers.europa.eu
metacareers.com/jobs
expert.ai/careers
careers.boozallen.com/jobs
activecampaign.com/about/careers/listing
amazon.jobs/*/jobs
jobs.lever.co
jobs.apple.com

Annex 2 – Full list of keywords used to identify degree types

bachelor
bachelor’s
bachelors
master’s
masters
master
bs
ba
ma
phd

Annex 3 – Sublime Text snippets instructions for corpus construction

Job title snippet:

```
<snippet>
<content><![CDATA[<section name="Job title">$SELECTION</section>]]></content>
</snippet>
```

Save the snippet as jobtitle.sublime-snippet in the Package directory of your Sublime Text installation. To create snippets for other sections you can use the above snippet as reference and modify the text inside quotes to create a new section for annotation.

To assign snippets to a key, you can modify the “User” key bindings .json file in Sublime Text by going to Preferences> Key bindings. Below you can find the .json file used for sections in WEBCTRL.

```
[
    { "keys": ["ctrl+1"], "command": "insert_snippet", "args": { "name": "Packages/User/jobtitle.sublime-snippet" } },
    { "keys": ["ctrl+2"], "command": "insert_snippet", "args": { "name": "Packages/User/keyinfo.sublime-snippet" } },
    { "keys": ["ctrl+3"], "command": "insert_snippet", "args": { "name": "Packages/User/jobdesc.sublime-snippet" } },
    { "keys": ["ctrl+4"], "command": "insert_snippet", "args": { "name": "Packages/User/jobfunctions.sublime-snippet" } },
    { "keys": ["ctrl+5"], "command": "insert_snippet", "args": { "name": "Packages/User/requiredqualifications.sublime-snippet" } },
    { "keys": ["ctrl+6"], "command": "insert_snippet", "args": { "name": "Packages/User/about.sublime-snippet" } },
    { "keys": ["ctrl+7"], "command": "insert_snippet", "args": { "name": "Packages/User/benefits.sublime-snippet" } }
]
```

References

Barbaresi, A. (2015). Ad hoc and general-purpose corpus construction from web sources. Linguistics. ENS Lyon. Retrieved from https://tel.archives-ouvertes.fr/tel-01167309

Barcaroli, G., Scannapieco, M., & Summa, D. (2016). On the use of internet as a data source for official statistics: A strategy for identifying enterprises on the web. Rivista italiana di economia, demografia e statistica, LXX(4), 25-41.

Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. International Conference on Language Resources and Evaluation.

Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. International Conference on Language Resources and Evaluation.

Bernardini, S., Baroni, M., & Evert, S. (2006). A WaCky introduction. In M. Baroni & S. Bernardini (Eds.), Wacky! Working papers on the web as Corpus (pp. 9–40). Bologna: GEDIT. Retrieved from: http://wackybook.sslmit.unibo.it/pdfs/bernardini.pdf.

Beveridge, A., Studies, W., & Gallagher, J. (2021). Project-Oriented Web Scraping in Technical Communication Research. Journal of Business and Technical Communication, 36. https://doi.org/10.1177/10506519211064619

Bowker, L. (2002). An empirical investigation of the terminology profession in Canada in the 21st century. Terminology, 8(2), 283–308. https://doi.org/10.1075/term.8.2.06bow

Bruthiaux P. 1996, The Discourse of Classified Advertising. Exploring the Nature of Linguistic Simplicity, Oxford University Press, New York.

Cignarella, A. T., Bosco, C., Patti, V., & Lai, M. (2018). TWITTIRÒ: an Italian Twitter Corpus with a Multi-layered Annotation for Irony. IJCoL, 4(2), 25-43.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990-2008+): Design, architecture, and linguistic insights. International Journal of Corpus Linguistics, 14(2), 159-190. https://doi.org/10.1075/ijcl.14.2.02dav

de Schryver, G. (2002). Web for/as corpus: A perspective for the African languages. Nordic Journal of African Studies, 11(2), 266–282.

Do Vale, J. W. S. P., Nunes, B., & de Carvalho, M. M. (2018). Project Managers' Competences: What Do Job Advertisements and the Academic Literature Say? Project Management Journal, 49(3), 82-97.

ELIS Research. (2023). European Language Industry Survey 2023: Trends, expectations and concerns of the European language industry. Retrieved from https://elis-survey.org/wp-content/uploads/2023/03/ELIS-2023-report.pdf

Ferraresi, A., Aragrande, G., Barrón-Cedeño, A., Bernardini, S., & Miličević Petrović, M. (2021). Competences, skills and tasks in today's jobs for linguists: Evidence from a corpus of job advertisements. Zenodo. https://doi.org/10.5281/zenodo.5030879

Fu, X. (2012), “The use of interactional metadiscourse in job postings”, Discourse Studies, Vol. 14 No. 4, pp. 399-417, doi: 10.1177/1461445612450373.

Garzone, G. E. (2018). Job advertisements on LinkedIn. Generic integrity and evolution. Lingue e Linguaggi, 26, 197-218.

Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. Journal of personality and social psychology, 101(1), 109–128. https://doi.org/10.1037/a0022530

Giampieri, P. (2021). An Analysis of the “Right of Termination”, “Right of Cancellation” and “Right of Withdrawal” in off-Premises and Distance Contracts According to EU Directives. Comparative Legilinguistics, 47, 105-133. https://doi.org/10.2478/cl-2021-0014

Gray, B., Egbert, J., & Biber, D. (2017). Exploring methods for evaluating corpus representativeness. Paper presented at the Corpus Linguistics International Conference 2017, University of Birmingham, 24-28 July.

Gundlapalli, A. V., Divita, G., Carter, M. E., Redd, A., Samore, M. H., Gupta, K., & Trautner, B. (2015). Taming Big Data: An Information Extraction Strategy for Large Clinical Text Corpora. Studies in health technology and informatics, 213, 175–178.

Harper, R. (2012). The collection and analysis of job advertisements: A review of research methodology. Library and Information Research, 36, 29-54. https://doi.org/10.29173/lirg499

Karacsony, P., Izsák, T., & Vasa, L. (2020). Attitudes of Z generations to job searching through social media. Economics and Sociology, 13(4), 227-240. doi:10.14254/2071-789X.2020/13-4/14

Kochetova, L. A., Sorokoletova, N. Y., Ilyinova, E. Y., & Volkova, O. S. (2017). Corpus-Assisted Comparative Study of British Job Advertisements: Sociocultural Perspective. In Proceedings of the 7th International Scientific and Practical Conference “Current issues of linguistics and didactics: The interdisciplinary approach in humanities” (CILDIAH 2017) (pp. 133-139). Atlantis Press. https://doi.org/10.2991/cildiah-17.2017.24

Kutter, A., & Kantner, C. (2012). Corpus-Based Content Analysis : A Method for Investigating News Coverage on War and Intervention.

Leech, G. (2007). New resources, or just better old ones? The holy grail of representativeness. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), Corpus linguistics and the web (pp. 133–149). Amsterdam: Rodopi.

Lipovac, I., & Marina Bagić Babac, M. (2021). Content analysis of job advertisements for identifying employability skills. Interdisciplinary Description of Complex Systems, 19(4), 511–525. https://doi.org/10.7906/indecs.19.4.5

Lopez-Mateo, C., & Olmo-Cazevieille, F. (2015). Compiling Texts for a Specialized Corpus in the Biochemistry Domain: Theoretical and Methodological Aspects. Procedia - Social and Behavioral Sciences, 198. https://doi.org/10.1016/j.sbspro.2015.07.448

Lotfi, C., Srinivasan, S., Ertz, M., & Latrous, I. (2021). Web Scraping Techniques and Applications: A Literature Review. doi:10.52458/978-93-91842-08-6-38

Ma, Y., Liu, Z., & Zhang, X. (2023). Adaptive Multi-Corpora Language Model Training for Speech Recognition. In ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). Rhodes Island, Greece. https://doi.org/10.1109/ICASSP49357.2023.10096577

Mang, C. (2012). Online job search and matching quality (ifo Working Paper Series No. 147). ifo Institute - Leibniz Institute for Economic Research at the University of Munich. https://EconPapers.repec.org/RePEc:ces:ifowps:_147

Maulan, S., Jailani, A. I., Ayob, N. M., Jing, H. F., & Yusuf, A. H. S. M. (2023). The language industry's insight into the curriculum design of a language programme. European Proceedings of Educational Sciences.

Matta, P., Sharma, S., & Uniyal, N. (2022). Comparative Study Of Various Scraping Tools: Pros And Cons. In 2022 IEEE Delhi Section Conference (DELCON) (pp. 1-5). doi:10.1109/DELCON54057.2022.9753358

McEnery, T. (2003). Corpus Linguistics. In R. Mitkov (Ed.), The Oxford Handbook of Computational Linguistics (pp. 448–463). Oxford University Press.

McEnery, T., & Hardie, A. (2011). Corpus linguistics: Method, theory and practice. Cambridge University Press.

McMillan, R. (2023, July 12). AI Junk Is Starting to Pollute the Internet. The Wall Street Journal. Retrieved December 1, 2023, from https://www.wsj.com/articles/chatgpt-already-floods-some-corners-of-the-internet-with-spam-its-just-the-beginning-9c86ea25

Paquot, M., & Gries, S. T. (Eds.). (2021). A practical handbook of corpus linguistics (1st ed.). Springer Nature.

Panta, D. (2015). Web crawling and scraping : developing a sale-based website. Turku University of Applied Sciences. https://urn.fi/URN:NBN:fi:amk-201505035716

Pomikálek, J. (2011). Removing boilerplate and duplicate content from web corpora (PhD Thesis). Masaryk University, Faculty of Informatics, Brno, Czech Republic

Redman, T., & Mathews, B. P. (1992). Advertising for Effective Managerial Recruitment. Journal of General Management, 18(2), 29-44. https://doi.org/10.1177/030630709201800203

Sakurai, K. & Okubo, Y. (2017) Job Seeker Trends 2016: Increasing Global Mobility. Boston Consulting Group and Recruit Works Institute. Retrieved from https://www.works-i.com/pdf/170202_jst2016_eng.pdf

Sinclair, J. (1982). Reflections on Computer Corpora in English Language Research. In S. Johansson (Ed.), Computer Corpora in English Language Research (pp. 1-6). Bergen: Norwegian Computing Centre for the Humanities.

Sinclair, J. (1996). Preliminary recommendations on Corpus Typology (Tech. Rep.). EAGLES – Expert Advisory Group on Language Engineering Standards. Retrieved from https://ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html

Sinclair, J. (2005). Corpus and text – Basic principles, and appendix: How to build a corpus. In M. Wynne (Ed.), Developing linguistic corpora: a guide to good practice. Oxford: Oxbow Books. Retrieved from https://users.ox.ac.uk/~martinw/dlc/chapter1.htm

Sodhi, M. S., & Son, B.-G. (2009). Content analysis of O.r. job advertisements to infer required skills. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.1640814

Tarat, S., Siritararatn, N., & Jaroongkhongdach, W. (2021). A methodological integration of corpus analysis and content analysis. In Proceedings of the International Conference: DRAL4 2021 (pp. 117-130). Kasetsart University.

Tessari, M. (2017). Creazione di un corpus specialistico e studio della fraseologia: l'esempio della crisi economica 2014-2016. Università di Bologna. Retrieved from https://amslaurea.unibo.it/id/eprint/15381

Tognini-Bonelli, E. (2001). Corpus linguistics at work. Amsterdam/Philadelphia: John Benjamins Publishing Co.

Van Hoye, G., & Lievens, F. (2005). Recruitment-related information sources and organizational attractiveness: Can something be done about negative publicity? International Journal of Selection and Assessment, 13, 179–187

Ward, J. [@jakezward]. (2023, November 24). We pulled off an SEO heist that stole 3.6M total traffic from a competitor. We got 489,509 traffic in October alone. Here's how we did it: [Image attached] [Post]. X. Retrieved from https://twitter.com/jakezward/status/1728032634037567509

Sitography

https://www.english-corpora.org//coca/

https://universaldependencies.org/

https://datadome.co/customers-stories/

https://pypi.org/project/jusText/

https://github.com/rsling/PyRex

https://corpus.tools/wiki/Onion

https://scrapy.org/

https://www.crummy.com/software/BeautifulSoup/

https://www.selenium.dev/

https://www.import.io/

https://www.dexi.io/

https://www.indeed.com

https://www.lexically.net/wordsmith/

https://monoconc.com/

https://upskillsproject.eu/

https://no.sketchengine.eu/