Co-funded by the
Erasmus+ Programme
of the European Union

Competences, skills and tasks in today's jobs for linguists: Evidence from a corpus of job advertisements

UPSKILLS Intellectual output 1.3

Compiled by:
Adriano Ferraresi[*], Gaia Aragrande[*], Alberto Barrón-Cedeño[*],
Silvia Bernardini[*], Maja Miličević Petrović[*]

[*] University of Bologna

UPSKILLS: UPgrading the SKIlls of Linguistics and Language Students

Erasmus+ Programme
Key Action 2: Cooperation for Innovation and the Exchange of Good Practices Action
KA203: Strategic Partnerships for Higher Education

Grant Agreement Number: 2020-1-MT01-KA203-074246

Document status: Final 20/05/2021
Dissemination level: Public

UPSKILLS Consortium:

&

with financial support from

https://doi.org/10.5281/zenodo.5030879

Table of contents

Executive Summary

The Corpus-based Analysis of Job Advertisements is part of the UPSKILLS needs analysis. Its objective is twofold. First, it aims to provide an overview of the **knowledge, skills and competences** mentioned in job posts targeting graduates in language-related degrees or professionals with expertise in this area, as well as of the typical **tasks and responsibilities** associated with these positions. Second, it aims to provide an initial list of companies at the crossroads between the language sector and the digital sector, which can be involved as stakeholders for the dissemination of UPSKILLS results.

Complementing this deliverable is a **richly annotated corpus of job ads** that has been made freely available to the research community. The corpus features slightly fewer than 200 job advertisements which describe positions requiring a combination of language/linguistics and digital or research skills. Despite its small size, the corpus is a highly curated one, with an overall balanced distribution across the three major sources of job ads, i.e. websites of technology companies, sections dedicated to job announcements in specialised websites dedicated to linguistics, and general-purpose job platforms.

The analysis was carried out adopting a corpus-driven, exploratory approach, identifying and classifying recurrent patterns on the basis of frequency lists extracted from the corpus. **Four main categories of skills and competences** emerge as salient, each of which can be associated with a **distinctive set of tasks** and responsibilities detailed in job ads. The first is **data and research skills**. Collecting and analysing data is mentioned both as an ability that job candidates should possess and as a task that they are expected to carry out. Analytical skills are also mentioned in relation to research tasks, whereby employees are required to conduct exploratory work focusing on language data, software tools, business processes and/or market needs, and produce written reports on their work. The second category is that of **technical skills**, which encompass knowledge of a programming language (usually Python), and methods deriving from computational linguistics or NLP. The area of application is that of software tools and machine learning models, which job candidates are expected to develop, test and improve. Thirdly, the requirement for **language and linguistics** disciplinary **knowledge** emerges clearly, as testified by the frequency with which linguistics (including semantics, syntax and morphology) and translation/localisation are mentioned. This category involves tasks such as translation and localisation, linguistic annotation and transcription of audio files.

Finally, **communication, interpersonal and organizational skills** are among the most frequently mentioned requirements. The high frequency of mentions of these skills might be related to the fact that they are crucial for tasks which are transversal across profiles and in high demand, such as working in teams, managing projects and interacting with clients and vendors.

List of abbreviations

Abbreviation	Definition
BA	Bachelor of Arts
BS	Bachelor of Science
CAT tools	Computer-Assisted Translation tools
MA	Master of Arts
MS	Master of Science
n-gram	Contiguous sequence of words of length n (e.g. 2-gram)
PhD	Doctoral degree
POS	Part-of-Speech
SD	Standard deviation
STEM	Science, technology, engineering, and mathematics

List of Figures

List of Tables

1. Rationale

The Corpus-based Analysis of Job Advertisements (also called job ads or job posts in this report) is part of the UPSKILLS needs analysis. Its objective is twofold. First, it aims to provide an overview of the knowledge, skills and competences required by the world of work on the basis of job posts targeting graduates in language-related degrees or professionals with expertise in this area, as well as typical tasks and responsibilities associated with these positions. Second, it aims to identify an initial list of companies at the crossroads between the language and digital sector, which can be involved as stakeholders for the dissemination of UPSKILLS results. The byproduct of this deliverable is a carefully controlled and richly annotated corpus of job ads that has been made freely available to the research community.

Together with the results of the questionnaire for companies and the focus groups with company representatives, the findings of the corpus-based analysis of job ads will contribute to defining the market's perspective on the expected profile(s) of professionals with language- and linguistics-related expertise, thus informing decisions on the interventions and materials to be created by the UPSKILLS team.

Section 2 reports on the procedure adopted to identify, sample and process relevant job ads, and describes the features of the resulting text corpus. Section 3 then presents the analysis of the corpus, which focuses on: a) qualifications (in terms of formal education and/or professional experience) that job candidates are expected to possess; b) the job functions that they are expected to fulfill; and c) the names typically associated with these jobs. Section 4 concludes by summarising the main findings.

2. The UPSKILLS corpus

2.1 Selection of job ads and corpus creation

The search for corpus texts started from the identification of websites publishing relevant job ads (criteria for text selection are listed below). Three main types of sources were considered:

- websites of technological companies (e.g. Amazon, Facebook); these were identified on the basis of the authors' personal knowledge and a listing of companies[1] which are known to hire linguists;

[1] https://careerlinguist.com/linguist-friendly-organizations/ (visited 28 April 2021)

- job ads posted to the LINGUIST List[2] or featured in the "Job postings" section of the Career Linguist website;[3]
- general-purpose employment platforms (e.g. LinkedIn[4]); only platforms which did not require paid registration were considered; in case considerable overlap was observed between the posts published by several platforms, only one post was retained, to avoid (near)duplication of information.

The search strategies varied according to the source considered. Manual perusal of job offers was performed for smaller company websites. In the case of general-purpose employment platforms and larger company websites which offered job search functionalities, a search was performed for the keywords: "linguist", "linguistics", "data", "language specialist". "NLP" was excluded as a keyword as it mostly returned engineering-oriented jobs, where a MS/PhD in STEM disciplines was usually required; while it also returned a few relevant job ads (e.g. for linguists with computational skills), a random check revealed that the same texts had been traced using "linguist" as a keyword. A combination of manual perusal and keyword-based searches was adopted to identify job offers from the LINGUIST List. The bulk of the searches were performed between December 2020 and January 2021. Only posts in English were considered. In view of the global/international nature of most of the companies surveyed, and the fact that jobs are increasingly open to remote applicants, no attempt was made to control for the countries where companies are based, or where jobs are offered.

The criteria for text selection were based on the job profile envisaged in the initial needs analysis conducted for the UPSKILLS project proposal. Job ads included in the corpus thus involve language- or linguistics-related tasks requiring digital and/or research skills. Job posts where a degree (MS or PhD) in a STEM field was a requirement, and jobs involving almost exclusively a) content creation tasks (e.g. writing/editorial jobs), or b) translation and revision tasks, were excluded.[5] Since the focus of the UPSKILLS project is on competences and skills required by private sector employers, academic jobs were also excluded. Job openings by institutions and national/transnational bodies were also not explicitly targeted, though they were not actively excluded either, as this would involve making decisions about the administrative (public/private) status of the employer. If multiple positions were advertised for candidates with different language competences (e.g. "Computational Linguist – Thai" and "Computational Linguist – Dutch"), only one version of the text was kept.

Selected job posts were manually converted into the TXT format, and the web page on which they were found was saved in PDF or HTML format to keep track of the original text

[2] https://linguistlist.org (visited 28 April 2021)

[3] https://careerlinguist.com/category/job-postings-2/ (visited 28 April 2021)

[4] https://www.linkedin.com/ (visited 28 April 2021)

[5] In several job ads the term "linguist" was used to refer to translator positions. Unless these positions involved skills or tasks other than translation (such as problem solving and/or technological ones), the texts were discarded, consistently with our general selection criteria.

outline. This is especially crucial for this type of text, which is extremely volatile: many job offers were no longer available online just a few weeks after they were collected, hence the need to save a local version (rather than a URL list).

In the final steps of the corpus construction process, TXT files were manually enriched with contextual metadata (e.g. original URL, source of job ad, and company), and annotated according to a pseudo-XML schema which was developed ad hoc to structure texts into several sections (e.g. job title, required qualifications, job functions; cf. Section 2.2.). Finally, texts were automatically annotated with part-of-speech (POS) and lemma information using the annotation tools provided by the SketchEngine platform,[6] and indexed for consultation with the NoSketch Engine toolset (Rychlý 2007; Kilgarriff et al. 2014)

2.2 Description of the UPSKILLS corpus

The UPSKILLS corpus is a corpus of job advertisements which target graduates in language-related degrees or professionals with expertise in language-related fields, and which describe positions requiring a combination of language/linguistics and digital or research skills. Table 1 reports information on corpus size and composition, and Figure 1 displays data on the provenance of texts.

Tokens	107,421
Number of texts	197
Average text length (+ SD)	544.1 (242.7)
Number of companies	112
Average number of texts per company (+ SD)	1.8 (2.1)

Table 1. Information on size and composition of the UPSKILLS corpus

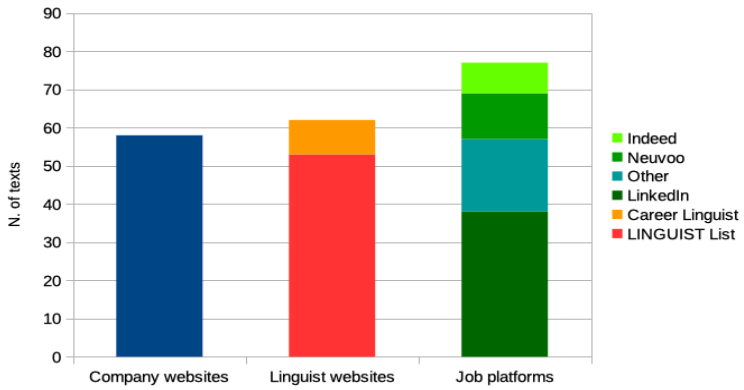

Figure 1. Sources of corpus texts

[6] https://www.sketchengine.eu (visited 28 April 2021)

The corpus is rather small in size (little over 100,000 tokens and fewer than 200 texts), yet it is a highly curated one, as text selection was performed completely manually, so as to guarantee relevance, consistency and coverage with respect to the UPSKILLS objectives. Text sources are varied, with an overall balanced distribution across the three major source types, i.e. websites of technology companies, sections dedicated to job announcements in specialised websites for linguists and general-purpose job platforms. The latter is the most represented category, and nearly half of the texts included there derive from the LinkedIn platform.

Each text in the corpus is annotated according to a pseudo-XML schema that makes it possible to store and represent:

- contextual metadata on text origin (URL, source where the text was published, and company offering the job) and on the text itself (text ID, name of the job offered, and link to an HTML/PDF copy of the text); this information is stored in the text header;
- structural information; on the basis of a manual analysis of a selection of ads, their prototypical structure was identified; the texts were then manually split to make it possible to carry out queries targeting only specific sections. The core component of the ads is enclosed in the job description (**jobdesc**) section tag, which may enclose two further sub-sections, one devoted to job functions which prospective employees are expected to carry out (**jobfunctions**), and one devoted to their required qualifications (**requiredqualifications**). If required qualifications and job functions were not presented separately in the job ads, only the **jobdesc** tag was assigned. The **jobdesc** and job title (**jobtitle**) sections were present in all the analysed ads. Additional sections appeared only in some of the texts: summaries of info on jobs (**keyinfo**), sections providing information on the company and company ethics (**about**) and sections on expected pay, perks, training opportunities and benefits for employees (**benefits**).

The structure of the text header and the full set of structural metadata are shown in Tables 2 and 3.

<text id="Actcamp001" url="https://www.activecampaign.com/about/..." source="Job platform – Other" job_name="Compliance Analyst" company="AM Team" file="http://corpora.dipintra.it/texts/upskills/Actcamp001.pdf">	id => File ID url => Original text URL source => Type of text source (company website / linguist website / job platform) job_name => Name of job as specified in the job ad company => Company publishing the offer file => PDF/HTML copy of text

Table 2. Example of Text header in the UPSKILLS corpus

<text id="", url="", source=""… >	Text header (cf. Table 2)
<section name="**jobtitle**">	Name of job
<section name="**keyinfo**">	Summary of info on job
<section name="**jobdesc**">	Core part of the ad
<section name="**jobfunctions**">	Job functions
<section name="**requiredqualifications**">	Required qualifications
<section name="**about**">	Info on company, including ethics
<section name="**benefits**">	Pay, perks, training opportunities and benefits

Table 3. Structure of texts in the UPSKILLS corpus

In addition to contextual and structural annotation, the texts are annotated with lemma and POS information, based on the default English tagset by the SketchEngine.[7]

The corpus is publicly available via the NoSketch Engine platform of the Department of Interpreting and Translation of the University of Bologna.[8]

3. Corpus-based analysis of job ads

3.1 Method of analysis

The corpus data were analyzed using Microsoft Excel and the NoSketch Engine platform tools. In line with the aims outlined in Section 1, targeted queries were created to identify salient phrases in each of the three main post sections, i.e. Required qualifications, Job functions, and Job title.

The procedure to extract and analyse corpus data was mostly exploratory, or corpus-driven (Tognini Bonelli 2001). An initial query was run to identify, section by section, the most frequent a) words/lemmas, b) sequences of words of length 2 to 4 (henceforth 2- to 4-grams),[9] and c) noun phrases containing a noun pre-modified by another noun or an adjective.[10] For the

[7] https://www.sketchengine.eu/english-treetagger-pipeline-2/ (visited 28 April 2021)

[8] https://corpora.dipintra.it/public/run.cgi/first_form?corpname=upskills;align= (visited 28 April 2021)

[9] E.g. `[tag!="SENT|SYM" & word="[A-Za-z\-]+"]{2,4} within <section name="Required Qualifications" />`; see https://www.sketchengine.eu/penn-treebank-tagset/ for the tagset used (visited 28 April 2021).

[10] E.g. `[tag="NN.*|JJ.*"] [tag="NN"] within <section name="Job functions" />`

most frequent phrases thus extracted, collocates were generated[11] and concordances checked. This made it possible to exclude results deriving from texts by single companies. The collocates were finally classified bottom-up to group required qualifications, job functions and job titles into meaningful, recurring categories.

3.2 Required (or preferred) qualifications

The bottom-up analysis of the most frequent 2- to 4-grams in the Required qualifications sections of job ads reveals a range of requirements that can be broadly classified into three main categories: a) formal education, b) experience and knowledge, and c) skills and abilities. One aspect that should be highlighted is that the qualifications described in this section are not necessarily compulsory; rather, they may be construed as preferred, as in example 1 (the string in square brackets refers to the corpus text ID):

1. 'Professional work experience strongly preferred but not required.' [Linguist030]

The various categories of qualifications are illustrated in detail in the following sections.

3.2.1 Formal education requirements

A Bachelor's degree is mentioned as a relevant qualification in less than half of the scrutinised job ads (40.6%) and the percentage is even lower for Master's degrees (35%) and Ph.D.'s (10.7%) (see Table 1). We do not make a distinction between Bachelor's or Master's degrees in humanities or social sciences (BA/MA) vs. scientific fields (BS/MS), as we excluded a priori job ads for which a degree in STEM fields was a requirement. It should also be noticed that these requirements are not mutually exclusive, as job ads might mention a Bachelor's or Master's degree as alternatives (example 2), nor in fact always compulsory, (example 3):

2. 'Minimum of Bachelor's Degree in Linguistics, Computational Linguistics or related disciplines' [Linguist009]
3. 'BS/BA Degree and/or MBA preferred' [Apple002]

Required degree	Number of texts	Percentage of sample
Bachelor's degree (BA or BS)	80	40.6
Master's degree (BS or MS)	69	35.0
Ph.D.	21	10.7

Table 4. Formal education requirements

[11] Settings: Range/span: -5 to 5 (i.e. 5 words to the right or the left of the searched word). Minimum frequency of collocates in corpus: 3. Minimum frequency of collocation as a whole (i.e. searched word + collocate): 3.

Table 5 lists the names of the disciplines occurring as part of the expression "Degree in", which should provide an indication of the names of degrees looked for by employers.

"Degree in"	Frequency (occurrences)
linguistics	44
computational linguistics	22
computer science	10
information science / information systems	4
speech science	3 (from a single company)
cognitive science, library science, data science, social science	< 3

Table 5. Degree names occurring in the Required qualifications section

3.2.2 Experience and knowledge

This category encompasses various subcategories of qualifications. On the one hand, "experience" is used to refer to practical experience that job candidates are expected to have gained by working as professionals in a certain field. The expressions "work / industry / equivalent / practical / relevant experience" are usually employed in this sense and occur in 57 job ads (28.9% of sample). This type of experience is sometimes mentioned as an alternative to formal education requirements, as in example 4:

4. 'Degree in Library Science, Information Systems, Linguistics or equivalent professional experience' [Amazon008]

A follow-up query reveals that the number of required years of work experience normally ranges from 2 to 8.

On the other hand, "experience" is used to refer to the set of domains, concepts and tools with which candidates are expected to be familiar. In this second sense, the word is used as a synonym for "knowledge" or "understanding". The analysis of collocates for these words, complemented by a query targeting the most frequent noun+noun or adjective+noun phrases, reveals 4 main areas of knowledge/experience: a) familiarity with data, tools and techniques; b) language competences, c) knowledge of (academic) disciplines, and d) other competences. These are illustrated in Table 5; only words/concepts mentioned in more than 5% of texts are shown and discussed.

Category	Specific knowledge / competences	Number of texts (% of sample)	Context phrases
Data, tools and techniques	data	68 (34.5)	language data; data structures; large datasets; large quantities of data; data analysis
	tools + software	62 (31.5) + 45 (22.8)	command line tools; marketing automation tools; CAT tools; transcription systems; Microsoft Office; version control systems
	programming	36 (18.3)	Python; programming language; scripting language
	annotation	29 (14.7)	Data annotation; data markup; linguistic annotation
	machine learning	28 (14.2)	-
	social media	11 (5.6)	-
	regular expressions	10 (5.1)	-
Academic disciplines	linguistics (excl. computational linguistics)	94 (47.7)	-
	computational linguistics	50 (25.4)	-
	Natural Language Processing / NLP	38 (19.3)	-
	semantics	36 (18.3)	-
	translation	34 (17.3)	-
	syntax	32 (16.2)	
	localization	22 (11.2)	-
	morphology	14 (7.1)	-
Language competences	languages (excl. programming/scripting l.)	50 (25.4)	multiple languages; foreign languages; target languages; second languages
Other	research	38 (19.3)	research design; quantitative research; qualitative research; market research
	project management	20 (10.1)	-

Table 6. Experience and knowledge requirements

The knowledge and experience that appear to be most in demand in the analysed job ads are related to data, tools and techniques. These overlap to a large extent with the categories of data acquisition and handling skills identified as being underrepresented in the UPSKILLS

Survey of Curricula. They include generic abilities to analyse or annotate (large quantities of) language data, as well as knowledge of specific tools or software (e.g. command-line tools or CAT tools) and techniques to analyse or manipulate data (e.g. machine learning or regular expressions). Among programming languages, Python is the one that is mentioned most frequently.

Familiarity with the concepts and methods of specific academic disciplines also features prominently in the Required qualifications section. Linguistics is the most frequently mentioned discipline, including the subfields of semantics, syntax and morphology. Knowledge of computational linguistics and Natural Language Processing is also a frequent requirement, followed by translation and localization.

Finally, 25% of jobs ads mention knowledge of at least one foreign language as required or preferred. Other required types of experience include those related to research activities (both quantitative and qualitative) and project management.

3.2.3 Other skills and abilities

The last category of requirements concerns abilities and skills. As can be expected, some overlap is observed between this macro-category and that of required experience and knowledge. The main difference lies in the fact that these abilities and skills are non-disciplinary. Table 7 summarises the abilities and skills that are mentioned in at least 5% of the texts.

Ability / Skill	Number of texts (% of sample)	Context phrases
communication skills	81 (41.1)	written communication skills; oral/verbal/spoken communication skills
attention to detail	59 (29.9)	strong/excellent attention to detail
organizational skills	36 (18.3)	good/proven organisational skills
analytical skills	25 (12.7)	-
problem solving (or problem-solving)	22 (11.2)	strong/excellent problem solving skills
interpersonal skills	18 (9.1)	-
fast-paced	17 (8.6)	ability to work/thrive in a fast-paced environment
ability to work independently	13 (6.6)	-

Table 7. Skills and abilities

Some of the skills mentioned in this category, i.e. analytical and problem solving skills overlap with those mentioned as “research skills” in the UPSKILLS proposal and the Survey of Curricula. Together with the category of “research experience” discussed in Section 3.1.2,

these testify to the importance of broadly defined research-related skills in the analysed job posts. Among other skills and abilities, relational ones seem prominent: written and spoken communication skills are mentioned in 41.1% of the sample, and interpersonal skills in 9.1%. Attention to detail (29.9%), organizational skills (18.3%) and the ability to work in a fast-paced environment (8.6%) and/or independently (6.6%) also occur frequently.

3.3 Job functions

The analysis of job functions was carried out by restricting queries to the Job functions section of the corpus. The starting point in this case was a list of the 50 most frequent lemmas, for which collocates and concordances were generated and classified bottom-up. Two main categories of job functions emerged, i.e. a) linguistic, research- and technology-focused tasks and b) general tasks. These are described in detail in the following sections.

3.3.1 Linguistic, research- and technology-focused tasks

This category includes the tasks that were expected to characterize the job ads included in the UPSKILLS corpus, i.e. tasks for which a combination of language-, research- and technology-related competences and skills are required (see Table 8).

Task	Noun(s) occurring in corpus	Number of texts (% of sample)	Context phrases
Quality assurance	Quality	77 (39.1)	perform quality controls/assurance; improve quality (of tools/data output)
Data acquisition and handling	Data, information, materials	71 (36.0)	analyse data; collect data; extract information; categorize materials
Tool development and use	Tools	56 (28.4)	improve or develop software/ technological/NLP tools; use internal tools
Research work	Research	41 (20.8)	conduct research; support or participate in research and development
Translation and localisation work	Translation, localisation	37 (18.8)	ensure consistency of localizations; provide translations
Model building and testing	Model	30 (15.2)	build language models; train models
Annotation	Annotation	23 (11.7)	perform data annotation; update/create annotation guidelines
Performance check and improvements	Performance	22 (11.2)	analyse, test or improve (system/product) performance
Transcription work	Transcription	15 (7.6)	perform phonetic transcription; provide transcription for audio clips

Table 8. Linguistic, research- and technology-focused tasks

The tasks presented in Table 8 can be roughly assigned to the three main categories mentioned in the title of this section. Quality assurance, which ranks first with its mentions in 39.1% of the sample, is perhaps an outsider to this classification, as it encompasses tasks such as performing quality checks or improving quality e.g. of NLP tools or data outputs; in this light, it is cross-cutting with respect to such categories. Linguistic tasks involve language- or linguistics-related competences, and include translation and localization (18.8% of texts), annotation work of language data (11.7%) and transcription of audio files (7.6%). Research tasks encompass data acquisition and handling jobs (36%), which usually consist in collecting and analysing/categorizing language data, and research work per se (e.g. participating in research and development; 20.8%). The remaining tasks are highly technological in nature, and consist in developing NLP or other tools (28.4), testing or improving their performance (11.2) and building or training language models in the context of machine learning (15.2%).

3.3.2 General tasks

This category encompasses tasks which are not specific to language- or technology-related jobs. These are summarised in Table 9.

Task	Noun(s) occurring in corpus	Number of texts (% of sample)	Context phrases
Teamwork	Team	111 (56.3)	work with or support teams (e.g. product team, engineering team, project team, development team); collaborate with team members
Project management	Project	89 (45.2)	manage, lead or oversee projects
Customer service or support	Customers, clients	78 (39.6)	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
Report writing	Report	37 (18.8)	provide written reports; write up project reports
Relation with vendors	Vendors	21 (10.7)	work with or support external vendors; assess vendors' performance

Table 9. Job functions: general tasks

The majority of ads involve teamwork as a component of the job (56.3%), followed by project management (45.2%) and customer service or support (39.6%). Interestingly, report

writing features in nearly 20% of the texts: this often relates to the ability to (analyse and) summarise data, which has been shown to be one of the main job requirements in Section 3.1.2. Examples of general task mentions are:

5. Maintain project trackers, monitor key project metrics and provide project reports as required [Appen002];
6. The Language Analyst will interpret/translate and provide reports on essential elements of information [LinkedIn017]

Relation with vendors, which can also be seen as a component of project management, is mentioned in 10.7% of the sample.

3.4 Job titles

Unlike the other corpus sections analysed in this report, the Job title section usually consists of a single phrase, which may be idiosyncratic to the job post being analysed, or to the company publishing them. The analysis could not therefore be based on the criteria adopted in Sections 3.1 and 3.2, where only phrases occurring in at least 5% of the texts were taken into consideration. In this case, we start by drawing a word cloud[12] based on a frequency list derived from the Job title section (only words occurring 3 or more times are included; Figure 2), followed by an analysis of the most frequent 2- to 3-grams (Table 10).

Figure 2. Word cloud based on the Job title section

[12] Generated using the online tool https://www.wordclouds.com (visited 28 April 2021)

Keyword	Job title	Frequency
Linguist	computational linguist	16
	associate linguist	11
	linguist	8
	analytical linguist	6
	data linguist	4
Data	data scientist	5
	data linguist[13]	4
	data analyst	3
Manager	project manager	11
	language manager	3
	localization project manager	3
Other	speech scientist	4
	language analyst	3
	project coordinator	3

Table 10. Most frequent job titles

The keyword "linguist" features prominently in the analysed job titles, both in terms of the variety of names in which it enters, and in terms of its frequency. It seems to refer to three main profiles, i.e. that of a) "computational linguist", which is the overall most frequent job title, b) "linguist" or "associate linguist", and c) "data" or "analytical" linguist, which specifies a data analysis component in the job name; in this light, the latter variants are probably synonyms for the name "language analyst". "Data" is also a recurring word in job titles, being featured in names such as "data scientist" and "data analyst". Finally, "manager" occurs as part of the frequently occurring title "project manager", a synonym of which seems to be "project coordinator"; the type of project or areas of work to be coordinated is sometimes specified, as in "localization project manager" and "language manager".

The degree to which these job titles correspond to consistent (sets of) job profiles remains an open question, which will be further investigated through the questionnaires and focus groups with companies carried out as part of the UPSKILLS project.

[13] Note that this title was also retrieved by looking at "linguist".

4. Summing up and concluding remarks

The corpus analysis of job ads aimed to survey and examine jobs which involve competences and skills at the crossroads between languages and linguistics, technology and research (excluding better established jobs such as translator or localizer). The corpus, which has been made available to the public, includes slightly fewer than 200 job ads, all of which underwent manual scrutiny both in the phase of text selection and in the phase of annotation with contextual and structural metadata. The analysis was carried out adopting a corpus-driven, exploratory approach, identifying and classifying recurrent patterns based on frequency lists extracted from the corpus.

On the basis of the UPSKILLS corpus, four main categories of skills and competences emerged as particularly salient,[14] each of which can be associated with a distinctive set of tasks and responsibilities detailed in job ads. The first one is the category of **data and research skills**. Collecting and analysing data, often qualified as language data, is both mentioned as an ability that job candidates should possess and as a task that they are expected to carry out. Analytical skills are also mentioned in relation to research skills, which seem to partly overlap with data skills. Employees are expected to conduct research focusing on data, tools, business processes and/or market needs, and produce written reports on their work – a requirement which also ties in with the need to have excellent written communication skills (cf. infra).

The second category is that of **technical skills**, which include specific instrumental competences. Required or preferred technical skills encompass knowledge of a programming language (frequently Python), and methods deriving from computational linguistics or NLP. The most clearly identifiable area of application is that of software tools and machine learning models, which job candidates are expected to develop, test and improve.

Thirdly, the requirement for **language and linguistics disciplinary knowledge** emerges prominently from job ads, as testified by the frequency with which linguistics (and the subfields of semantics, syntax and morphology) and translation/localisation are mentioned in the Required qualifications section. This category involves tasks which are more linguistic than technological in nature, and include translation and localisation, linguistic annotation and transcription of audio files.

Finally, **communication, interpersonal and organizational skills** are among the most frequently mentioned requirements. For example, strong written and verbal communication skills are a transversal requirement featured in more than half of the examined profiles. However, the high frequency of mention of these skills might also be related to the fact that they are crucial for tasks which are themselves transversal and in high demand, such as working in teams, managing projects and interacting positively with clients and vendors.

One of the most striking aspects emerging from the analysis of job ads is that university degrees feature among required qualifications in less than half of the sample: mentions of a

[14] The terminology adopted here reflects the one proposed in the UPSKILLS literature review (Bernardini & Miličević Petrović 2021).

BA/BS degree appear in 40.6% of the sample, and of an MA/MS degree in 35% of the sample. By combining mentions of either a first-cycle or a second-cycle degree, results are still below 60% (58.9%, corresponding to 116 job ads). Even in these cases, the degree might constitute a preferred rather than a required qualification, and can thus be seen as an alternative to equivalent professional experience.

Data such as these seem to confirm issues which also emerged in the UPSKILLS Literature review and Survey of curricula. Language and linguistics degrees are experiencing a "branding problem" (Kelly, 2013: 8) whereby they are unable to sell themselves in the labour market, or, even worse, they might not be catering for what the market expects from them. The analysis of the job profiles carried out here suggests that competences related to data collection and analysis, research work, computational methods and strong interpersonal and organizational skills are part of the expected skillset of 21st century language and linguistics graduates.

References

Bernardini S., Miličević Petrović M. (2021). Toward a new profile for twenty-first century language specialists: Industry, institutional and academic insights. UPSKILLS Task Report.

Kelly, M. (2013). The Future of Language Degrees. Online: http://www.celelc.org/archive/Projects-and-Reports/999-Future-Language-Degrees/Languages_Degrees_final_report.pdf (visited: 24.01.2021).

Kilgarriff, A., V. Baisa, J. Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý and Vít Suchomel (2014). The Sketch Engine: ten years on. *Lexicography* 1: 7-36.

Rychlý, P. (2007). Manatee/Bonito: A modular corpus manager. In *Proceedings of the 1st workshop on recent advances in Slavonic Natural Language Processing*, 65–70. Brno: Masaryk University.

Tognini Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins.