# Which Domains, Tasks and Languages are in the Focus of NLP Research on the Languages of Europe?

**Diego Alves[1], Marko Tadić[1], Georg Rehm[2]**
[1] Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
[2]Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin, Germany
diego.alves@uni-saarland.de, marko.tadic@ffzg.hr, georg.rehm@dfki.de

**Abstract**

This article provides a thorough mapping of NLP and Language Technology research on 39 European languages onto 46 domains. Our analysis is based on almost 50,000 papers published between 2010 and October 2022 in the ACL Anthology. We use a dictionary-based approach to identify 1) languages, 2) domains, and 3) NLP tasks in these papers; the dictionary-based method using exact terms has a precision value of 0.81. Moreover, we identify common mistakes which can be useful to fine-tune the methodology for future work. While we are only able to highlight selected results in this submitted version, the final paper will contain detailed analyses and charts on a per-language basis. We hope that this study can contribute to digital language equality in Europe by providing information to the academic and industrial research community about the opportunities for novel LT/NLP research.

**Keywords:** European languages, language equality, language technology

## 1. Introduction

The fields of Natural Language Processing (NLP) and Computational Linguistics (CL) cover a wide range of topics. While CL draws from linguistics and NLP focuses more on computational methods, the terms are often used interchangeably. Language Technology (LT) is a neutral term encompassing both (Agerri et al., 2021). Today, Language Technology is integrated into various aspects of life. Recent progress has been driven by deep-learning models (Otter et al., 2020). Despite these advancements, challenges persist in achieving language equality, as outlined by a recent European Parliament (2018) resolution.

As the performance of machine learning and deep learning methods usually relies on large amounts of data, languages with smaller numbers of speakers are usually disadvantaged and endangered by digital extinction. With regard to Europe, the discrepancy regarding the availability of LT is highlighted by the reports of the European Language Equality (ELE) project describing the current status and challenges regarding LT for 39 European languages (Rehm and Way, 2023).

To promote digital language equality, it is crucial to understand individual language needs and by detecting their spot on the map of the NLP landscape. While initiatives like the European Language Grid (ELG, Rehm, 2023) contribute to the deployment of existing LT, it is also important to identify existing gaps concerning availability of resources designed for low-resourced languages.

We carried out a systematic analysis of current NLP research on Europe's languages with a specific emphasis on domains and NLP tasks. We analysed approx. 50,000 papers published in the ACL Anthology[1] between January 2010 and October 2022. Within this body of research, we identified the language, domain and NLP task a paper reports upon. One motivation behind this landscaping type of research was to identify popular domains and tasks as well as those that are very much under-researched. These gaps could potentially provided opportunities for novel research in the future. Our results provide a general overview into how NLP tools are used in different domains concerning Europe's languages and can be used by researchers to identify opportunities for future developments to promote language equality.

The remainder of this paper is structured as follows. First, Section 2 presents related work. Section 3 describes the methodology for information extraction based on a dictionary-based approach. Section 4 presents an evaluation of the dictionary-based approach and Section 5 highlights the general results regarding NLP tasks, domains, and languages. Section 6 describes a high-level overview of the results regarding the use of NLP tasks in different domains on a per-language basis. Section 7 concludes the paper.

## 2. Related Work

Current LT literature discusses technologies rather than domain-specific applications. Research papers describe new tools, methods and approaches and handbooks such as Mitkov (2022) provide

[1]https://www.aclweb.org/portal/



*Proceedings of the Second TDLE Workshop @LREC-COLING-2024*, pages 18–32

an overview of existing areas and resources. Although presenting important findings regarding the status of LT for different languages, surveys such as the ones presented in the language reports of the ELE project (Rehm and Way, 2023) or the META-NET White Papers (Rehm and Uszkoreit, 2012) do not present detailed analyses how tools are deployed in different domains.

A few articles describe the use of LT in specific fields. For example, Osterrieder (2023) present a complete overview of LT in finance. Additionally, several research papers present tools and resources for a particular domain, for example, a chemical tagger (Hawizy et al., 2011).

In Web of Science[2] and Scopus,[3] users can filter for specific domains, making it possible to find NLP articles in these domains. However, it is impossible to generate a complete overview.

This article presents a detailed analysis, using a dictionary-based approach, of the development of LT by the NLP community concerning different domains and languages with a focus on 39 European languages. The analysis is based on the ACL Anthology, which is why research published elsewhere (or not at all) is excluded. We are aware of the fact that supervised machine learning outperforms dictionary-based classification (Kroon et al., 2022), which is our approach, however, due to the large number of domains, tasks and languages, and because of the lack of annotated data to train models regarding this specific task, we decided to use the dictionary-based approach to establish this groundwork that can be the base for more advanced studies in the future. Our work is based on the EuLTDom project report[4] with evaluation results regarding the dictionary based approach and further analysis.

## 3. Data and Methodology

The ACL Anthology is an important Open Access archive with Open Source components for the NLP community. It is the main source of CL and NLP scientific literature and offers both text and faceted search features of the indexed papers and also author-specific pages. It allows open access to the proceedings of all ACL-sponsored conferences and journal articles, also hosting literature from sister organisations and their national venues (Gildea et al., 2018).

We used the ACL Anthology Corpus repository (Rohatgi, 2022) which provides PDF files, full-text, references, and other details extracted from the PDF files using GROBID.[5] This repository contains 80,013 articles and posters from 1957 to October 2022. We analyse a subset of this data, a total of 49,466 articles published between January 2010 and October 2022.

To understand the use of LT in different domains for different languages, we implemented a dictionary-based approach. We count the number of research papers in the subset of the ACL Anthology Corpus (see above) that mention the defined terms concerning languages, domains, and NLP tasks at least twice. In the first step of the analysis we look at each of these three dimensions separately, while in the second step, we count the number of articles that mention the domain/language/NLP task triple to identify how different domains use specific LT for each language. The lists of languages, domains, and NLP tasks to be used in the dictionary-based approach were defined in a way to avoid certain possible biases and are described in the following subsections.

### 3.1. Languages

We analyse the texts of papers written in English from the ACL Anthology for those 39 European languages for which an ELE Language Report exists.[6] For the languages that have more than one name (i. e., Catalan/Valencian and Romanian/Moldavian/Moldovan), while searching for the number of mentions in each paper, all possible names were considered. The complete list of languages is presented in Appendix A.

### 3.2. Domains

The list of relevant domains was defined following the Fields of Research and Development classification (FORD), which is the basis of the Frascati Manual (2015). This approach is closely related to and consistent with UNESCO's Recommendation concerning the International Standardisation of Statistics on Science and Technology (Unesco, 1978). The FORD classification provides a more complete set of domains when compared with the list considered in the ELE language reports (e. g., Melero et al., 2022). Although similar, the ELE list is shorter and includes general terms such as "Technology", "Science", and "Innovation".

We customised the FORD classification as follows: 1. the list was completed with ELE fields not present in the FORD one, excluding generic terms previously mentioned; 2. the FORD elements that correspond to the label "Other" (e. g., "Other natural sciences") were excluded; 3. the Health and Media domains were excluded because they were the focus of a concurrent study; and 4. terms such

[2] https://www.webofknowledge.com
[3] https://www.scopus.com
[4] https://tinyurl.com/356xt6b5
[5] https://github.com/kermitt2/grobid
[6] https://european-language-equality.eu

as “Economic geography” and “Social Geography” were replaced with “Geography”.

Our final classification contains 46 domains, which are clustered into five broader classes as presented in Appendix B.

### 3.3. NLP Tasks

The list of NLP tasks includes the information provided by Mitkov (2022) complemented with tasks found in the Wikipedia article on NLP[7] and two other tasks mentioned on the IBM website[8]: “Spam detection” and “Virtual agents and chatbots”. While Mitkov (2022) divides NLP/LT into two classes (i. e., tasks and applications), Wikipedia has a more detailed classification. The complete list contains 51 tasks divided into seven classes and is fully displayed in Appendix C.

### 3.4. Text Processing

We first attempted to analyse each of the three dimensions separately and analysed if every term listed above appeared at least twice in each individual article within the collection, using the Python Regular Expression operations library[9]. Preliminary tests, with a qualitative evaluation, showed that the main differences in the overall comparison of the languages, NLP tasks, and domain did not change using the threshold of two, five, or 10 occurrences. However, the total number of articles classified according to them is reduced when the threshold was increased. Thus, to improve the recall, we decided to keep the rule of minimum of two occurrences per article.

Texts and query terms were converted to lowercase for uniformity and, for each text available in the ACL Anthology Corpus, its full text (i. e., from abstract to conclusion) was analysed. The idea of considering only those articles where each term is mentioned at least twice is due to the fact that a certain term may be mentioned in the article even if the text is not exactly focusing on this term specifically but only mentioning it in passing.

Our goal was to examine how the NLP community has developed LT for different domains and languages. Most articles describe tools and other resources, thus the main topic here are neither the languages nor the domains. An article or poster is relevant for a certain language and domain if it clearly describes a concrete resource or application of an NLP task.

We also examined languages separately. First, the articles were analysed to check if a language was mentioned at least twice. Then, we checked if the article mentioned each domain/NLP task pair. With these results, heat maps were generated using the statistical data visualization Python library Seaborn.[10] The query concerning domains and NLP tasks was performed with the identified terms including synonyms and alternative orthographic forms. Special attention was required for some terms in the list of domains that may be used in different contexts, not necessarily referring to the domain, for example, “literature” and “history”. In these cases, besides the noun, the respective adjective also had to be mentioned at least once for the article to be counted (e. g., literature and literary; history and historical). Special treatment also had to be implemented for the domain “Arts” (or “Art”). As many papers contain the term “state-of-the-art” or its variations, a way to verify the context of the regular expression match was implemented to guarantee that these phrases are not counted.

The code and the results are available in the project’s GitHub repository.[11]

## 4. Evaluation

The dictionary approach relies on counting the occurrences of specific terms. This approach has inherent weaknesses when compared to methods for topic classification based on supervised machine learning and embeddings (Kroon et al., 2022). Considering the lack of explicitly annotated training data as well as overall resource restrictions, we opted for the keyword-based approach. To validate the efficiency of the dictionary-based approach, we decided to conduct an evaluation focusing on its precision.

In total, 49,466 articles from the ACL Anthology Corpus were analysed. In order to have a result with a confidence level of 95% and a 5% margin of error, the set to be analysed for the evaluation must contain a minimum of 382 articles.[12] As three dimensions are examined, we decided to select a sample of 130 texts for each one, a total of 390.

We randomly selected texts from the categorised ones, guaranteeing that the evaluation data has at least two representative texts for each of the terms considered as matches.[13] Furthermore, we verified that the articles cover all the years of the ACL Anthology we looked at (January 2010 to October 2022).

For each article of the evaluation data, we checked if the term found in the article really corresponded to a language, domain or NLP task name.

[7] https://en.wikipedia.org/wiki/Natural_language_processing

[8] https://www.ibm.com/topics/natural-language-processing

[9] https://docs.python.org/3/library/re.html

[10] https://seaborn.pydata.org

[11] https://github.com/dfvalio/EuLTDom2023

[12] Value determined using Calculator.net.

[13] Those terms contained in the lists that could not be found in the data set were omitted in the evaluation.

We considered it as a true positive if the term was used in the context of a resource (i. e., tool, model, data set, etc.) or a real application (e. g., evaluation of existing tools, surveys, etc.). False positives corresponded to the cases where the term was used in the context of a future research direction or for incorrect matches due to problems with regular expressions. Table 1 presents the results for each dimension and the overall precision.

| | Precision |
|---|---|
| Languages | 0.86 |
| Domains | 0.74 |
| NLP tasks | 0.84 |
| Overall | 0.81 |

Table 1: Precision for each dimension and overall

With regard to our overall objective, we consider the results of the dictionary-based method satisfactory. In comparison with the analysis conducted by Kroon et al. (2022), our results are comparable to the best machine-learning techniques. The domain dimension is the most problematic one, with a precision of less than 0.75. Below we present a qualitative analysis of the encountered errors.

We would like to stress that only precision was considered in this evaluation. It does not provide information on articles that present contributions regarding the three defined dimensions using different terms than the one contained in the lists. It seems plausible to imagine that the domain dimension should be the one with the lowest recall as the text may describe an application in a certain domain using a different name.

### 4.1. Languages

The errors observed when languages were analysed correspond mainly to the sections of the papers that deal with related or future work (44.1% of the errors). We also encountered other types of false positives: 1. The language is present in the name of an Organisation (e. g., "Norwegian University of Science and Technology"); 2. the language is mentioned in the context of a translation; 3. the term is mentioned as being excluded from a study; and 4. the term refers to a nationality, not the language itself.

From all the terms used in the regular expressions, only "Romani" was problematic as it was considered a match with words such as "Romanian" and "Romanized". Thus, for this specific language, our results should be handled with care.

We did not consider abbreviations or language codes such as ISO 639-3. Thus, if a language is only mentioned using its name once and then using an abbreviation, it was not counted as a match.

### 4.2. Domains

Regarding domains, the most frequent error corresponds to using the term in example sentences (32.2% of the false positives), e. g., "Civil engineering" (presented as an example of a compound).

The other most common error (31.3%) is related to the mention of the term in organisation names. It is present mostly in the Acknowledgement section or in the main text when departments are referred to. The term "Government" was specifically problematic as it was mentioned in copyright-related parts of certain articles (e. g., "The U.S. Government"). Besides "Government", two other terms created errors repeatedly: "History" and "Arts". The first one was sometimes used in contexts such as "history-dependent", the second one was considered a match with words such as "parts" and "parts-of-speech".

The analysis of false positives concerning the three dimensions shows that some errors are recurrent and, thus, can be easily corrected. In the case of terms appearing in related or future work, a condition can be established to guarantee that the term should not be considered if it appears only in these specific sections. Concerning problematic terms, more precise rules could also be defined to exclude erroneous matches.

### 4.3. NLP Tasks

For NLP tasks, most false positives were linked to using the terms in related or future work sections (57.3%). In a few cases, the term was mentioned as a task that was, however, not used in the paper, for example, when it is proposed as an alternative way to process the data.

Regarding problematic terms, we encountered errors relative to the acronyms "OCR" and "QA". The first was identified in words such as "democratic", and the second in Arabic words written with the Latin script (e. g., "qarAr" and "qAmato"). Moreover, "parsing" is the term used for constituency and dependency parsing. Nevertheless, when used in this analysis, it also matches with "Semantic Parsing" which is a specific term in the list of the NLP tasks.

## 5. Results

Next up, we present overviews of the three separate analyses concerning languages, domains, and NLP tasks.

### 5.1. Languages

Nearly all languages were found in the data set, the only language which does not appear in our data is Tornedalian. Of the 49,466 texts in the ACL

Anthology Corpus (from 2010 to 2022), 45,737 (92.5%) mention at least twice one of the languages from our set. However, they are not distributed homogeneously (see Figure 1), mirroring the findings of others (Gaspari et al., 2023) and also indicating a strong digital language inequality.

As expected, the most mentioned language is English (i. e., more than 20,000 articles), followed by German, French, and Spanish (more than 3,000 articles each). These results are compatible with similar studies such as Joshi et al. (2020) who present an analysis in terms of entropy of the LT disparity between languages using an older version of the ACL Anthology. Italian, Czech, and Portuguese have 1,000 to 1,500 articles each, and the vast majority of languages are mentioned in a number of articles between 100 to 1,000. Languages with this level of development benefit from existing resources to improve the status of their technologies by adapting tools already available for more resourced languages.

The languages with the smallest representation in our data set (less than 100 articles each) are Galician, Welsh, Maltese, Bosnian, Faroese, Saami, Karelian, Yiddish, Luxembourgish, and Tornedalian. These languages seem to be the most endangered ones regarding digital language extinction, thus requiring more attention from the NLP community.

These are general numbers concerning the ACL Anthology. (Joshi et al., 2020) show that conferences such as LREC tend to have more linguistic diversity than others. The dominance of English is also favoured by the fact that, usually, NLP resources are developed for this language and then deployed to others, thus, English results are also presented as a baseline.

We do not consider conferences that are not part of the ACL Anthology. Thus, the bigger picture that emerges out of this survey does not correspond precisely to the LT reality of each language. For example, the ACL Anthology does not include the proceedings of the Baltic HLT conferences, which focus on the Baltic languages.

## 5.2. Domains

Only 6,179 ACL papers (12.5% of the total) explicitly mention at least one of the terms from the list of domains. This may be explained by the fact that the focus of many articles is on the development of the tools and resources themselves, and not on their applications in specific areas. Furthermore, it is possible that some papers may have certain domains in mind but not refer to them explicitly. The complete list of domains and the respective number of mentions is presented in Appendix D.

"Linguistics" is the most cited domain which is an expected result as our data concerns work published in Computational Linguistics conferences. However, "Computer Science" is not so prominent, even though ACL papers also deal with this domain. The top ten most mentioned terms are from the Social Sciences and Humanities and the arts (varying from 2,783 to 351 articles). The first domain from a different class is Biological sciences, followed by other Natural Sciences domains such as "Physics", "Chemistry", and "Mathematics". Engineering and technology is the class of domains with the least number of articles.

Figure 2 shows the distribution of certain classes. Social sciences and Humanities and the arts correspond to 89.9% of the mentions. The dominance of the class Humanities and the arts is partially explained by the elevated number of mentions of the term "Linguistics" and the bias identified in the search for the term "Arts" in the texts.

The following domains were never mentioned: "Agricultural biotechnology", "Veterinary", "Animal and dairy science", "Industrial biotechnology", "Environmental Biotechnology", "Environmental engineering", "Materials engineering", and "Electronic engineering". This does not mean that LT is not used in these areas but it indicates that LT is not primarily developed specifically for them. Some of these terms are more specific than others, such as "Industrial biotechnology", "Environmental Biotechnology", and "Environmental engineering", therefore, it is possible that papers dealing with them may use other terms in the text.

A more thorough understanding of the current use of LT in Natural Sciences, Engineering and Technology, and Agricultural and Veterinary sciences is necessary for the identification of new opportunities in terms of more directed NLP development for these fields.

## 5.3. NLP Tasks

In total, 32,154 (65.0%) articles mention one of the NLP tasks at least twice. This percentage is higher than the one for domains but smaller than the one for languages. One reason for this may be that the coverage of our list is not sufficient.

We can observe that Machine Translation has been one of the main areas of the NLP community between 2010 and 2022. The number of articles mentioning MT is approximately twice as high as the number concerning the second most frequently mentioned task (Parsing). Question answering is ranked third.

The term "Parsing" encompasses many NLP tasks, thus, it may explain this higher rank. Furthermore, we can observe that tasks that are not higher-level NLP applications such as parsing, word segmentation, part-of-speech tagging, and named-entity recognition are positioned in the top ten of the most frequent ones. This can be due to

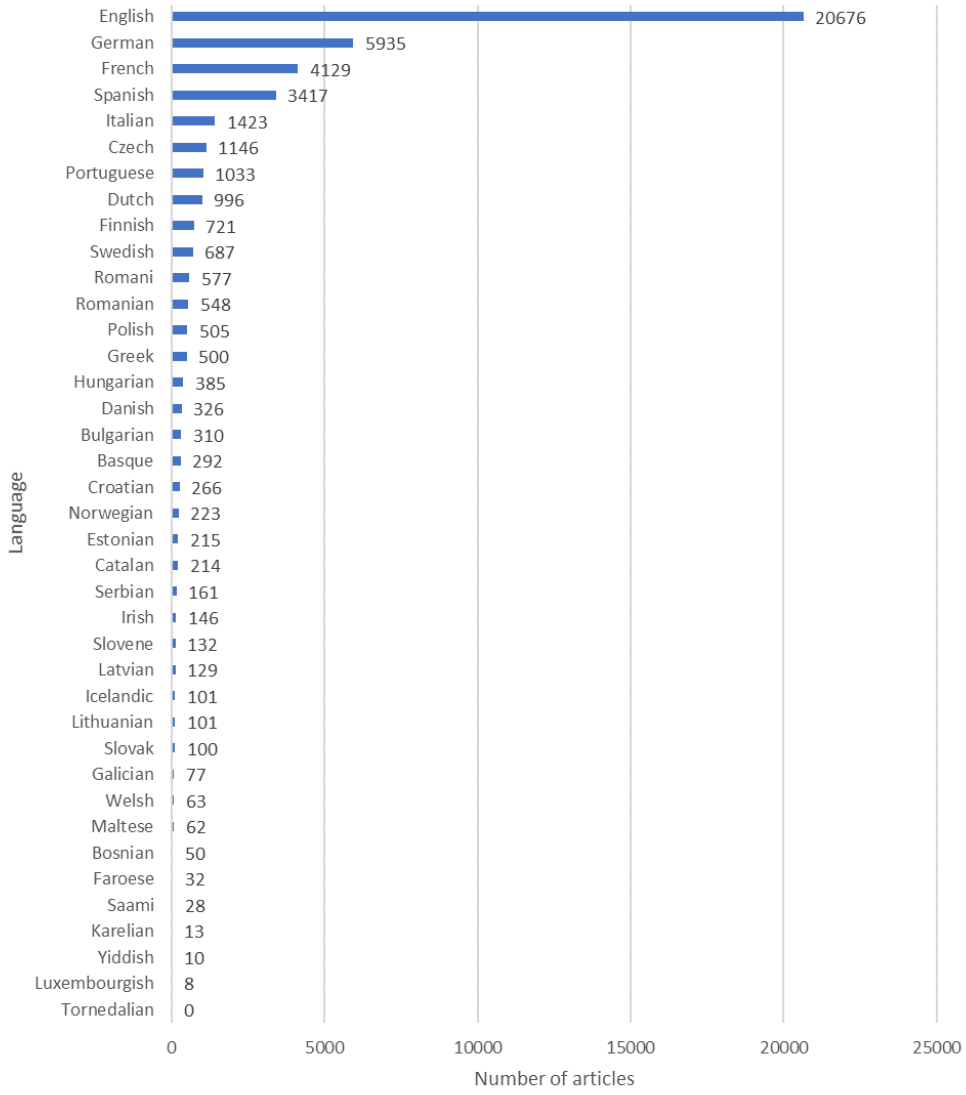


Figure 1: Mentions of European languages in the ACL Anthology (2010 until October 2022).

the fact that these tasks are part of more complex LT, being integrated into pipelines.

Of the 51 NLP tasks on our list, 39 (76.5%) are mentioned in less than 1,000 articles, thus, presenting a lot of potential for further development, e. g., deployment of existing architectures for languages other than English. Figure 3 presents the distribution of the NLP task classes. Almost half of the mentions correspond to higher-level NLP applications, due mostly to Machine Translation and Question answering.

Tasks with the lowest number of articles correspond to rather vague or very specific terms such as “Document AI” or “Implicit semantic role labeling”. The list of NLP tasks and the respective number of mentions is displayed in Appendix E. It would be useful to check if other names for these tasks are currently used by the NLP community to arrive at a more realistic view.

## 6. Results per Language

We present a detailed analysis concerning the use of LT in different domains per language (i. e., the number of articles where both domain and NLP task are mentioned at least twice each). The heat maps (x-axis: NLP tasks, y-axis: domains) provide a clear snapshot for each European language, and which can also be used as the basis of comparisons. All heat maps are available in the project’s GitHub repository.[14]

As expected, the languages with more mentions in the ACL Anthology result in more complete heat maps when compared to the languages with less mentions. However, we can clearly observe that not all domains and NLP tasks are not covered in recent research. Figure 4 shows the discrepancy in terms of technologies (i. e., data and tools) for different languages. Maltese is only mentioned in

[14] https://github.com/dfvalio/EuLTDom2023

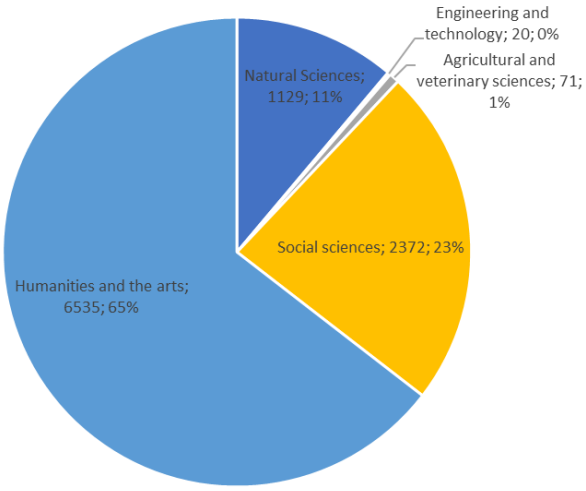


Figure 2: Number of articles presenting research about a certain class of domain.

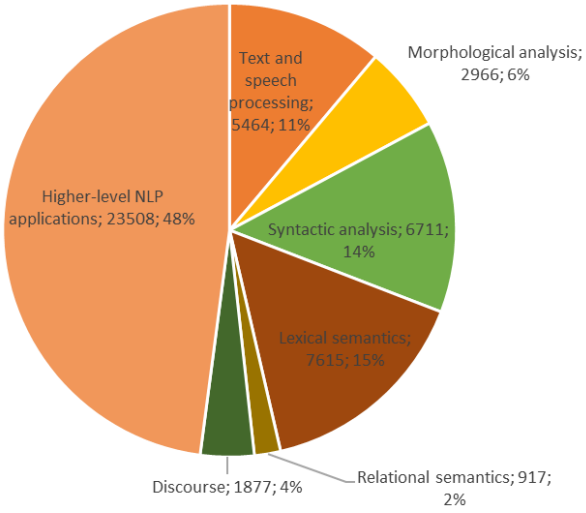


Figure 3: Number of articles presenting research about a certain class of NLP task.

62 articles, thus, its heat map is quite empty. On the other hand, for German (with 5,935 articles), the situation is much better, although still not comparable to the status of the NLP/LT development for English (with 20,676 articles).

Especially the gaps in the heat map of the English and other well-resourced languages can be used to identify new opportunities for the deployment of existing tools and algorithms. Furthermore, it is also possible to check what has been developed for closely related languages, which may facilitate cross-lingual transfer. We also generated a heat map with the overall use of NLP tasks by domains considering all European languages. As expected, "Linguistics" is the domain that has the highest number of associated NLP tasks.

Domains with relatively high usage of different types of LT (i. e., 20 articles or more) are "Arts", "Biological sciences", "Business", "Computer science", "Education", "Ethics", "Finance", "Government", "History", "Law", "Literature", "Physics", "Psychology", "Religion", "Sociology", and "Tourism". On the other hand, some domains use only specific NLP tasks. This is the case for "Ethics" with a predominance of articles on "sentiment analysis", "machine translation", and "question answering".

When examining the analysis regarding domains (except for Linguistics and Computer Science) that are most commonly associated with the top 10 tasks (i. e., tasks with at least 20 articles) we notice many similarities: "Business" and "Education" seem to be the domains that use most of the top 10 tasks. In Appendix F, we present these results in detail. The existence of more than 20 articles describing the use of LT in a specific domain seems to indicate that the specific application is well-developed and, thus, could represent an opportunity for low-resourced languages.

When we focus on the languages with less than 100 articles (excluding Tornedalian which was never mentioned), although the heat maps are very poorly populated, we can identify a few domains and tasks with at least some development. The "Business" and "Education" domains are usually associated with "Machine Translation" and "Natural Language understanding". "Education" is also sometimes mentioned in studies regarding "OCR", "part-of-speech tagging", and "speech-recognition". On the other hand, "History" is often associated with "speech recognition", "named-entity recognition", "machine translation", and "OCR". "Government" appears in association with "Natural Language understanding", "speech recognition", "question answering", and "machine translation", and "Biological Sciences" is usually associated with "information extraction", "named-entity recognition", "parsing", "machine translation", and "question answering".

Thus, it would be useful to check how the NLP data that was used in these papers can be applied to other tasks and deployed in other domains.

## 7. Conclusions and Future Work

We presented a mapping of NLP and Language Technology research onto 39 European languages and onto 46 domains. The analysis is based on almost 50,000 papers published between January 2010 and October 2022 in the ACL Anthology. The dictionary-based approach we use presents a satisfactory value of precision (i. e., higher than 0.80) when applied to identify how languages, domains, and NLP tasks are mentioned in articles contained in the ACL Anthology. We hope that this study can contribute to digital language equality in Europe by providing valuable information to the academic

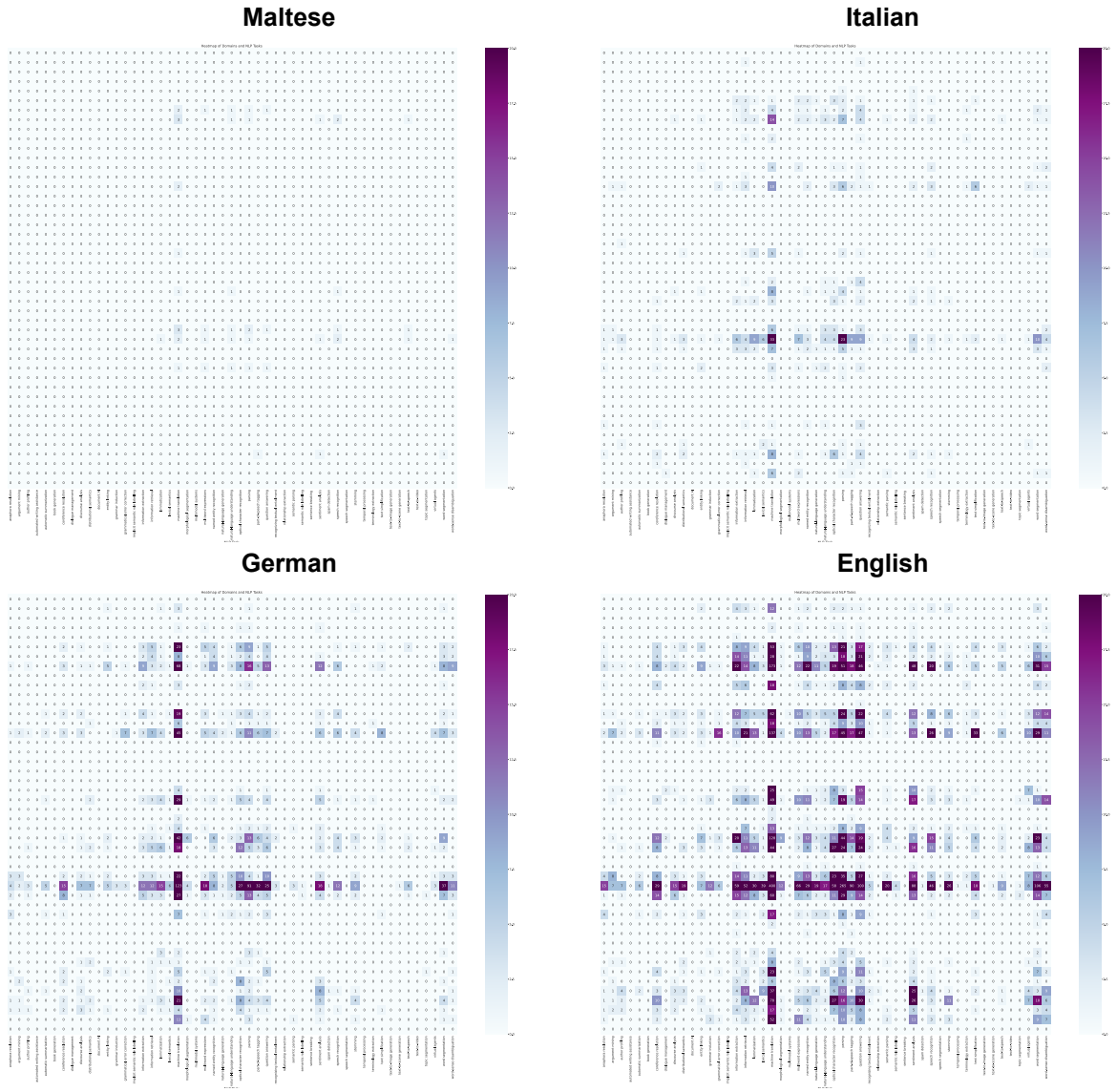


Figure 4: Comparison of four heat maps (Maltese, Italian, German, English).

and industrial research community about the opportunities for novel LT/NLP research.

This study only considers research published in the ACL Anthology. As a potential avenue for future work, a complementary study could be conducted considering other repositories such as Web of Science or Scopus, perhaps also fully structured repositories such as research knowledge graphs but these are too sparsely populated yet. Moreover, as ACL documents are only written in English, it would be useful to complete the analysis with the examination of papers written in the other listed languages. Furthermore, regular updates can be envisioned, for example, with new terms.

## 8. Ethics Statement

We affirm our commitment to conducting ethical research. We have followed established ethical guidelines and considered the broader societal implications of our work throughout the research process. We also respect copyright laws and intellectual property rights, giving proper attribution to the works of others in our research.

## 9. Acknowledgements

The European Language Equality project has received funding from the European Union under the grant agreements no. LC-01641480–101018166 (ELE) and no. LC-01884166–101075356 (ELE 2). The research presented in this article received funding from the project ELE 2 through the FSTP funding scheme under the title “European LT Domains 2023 (EuLTDom2023)”.

## 10. Bibliographical References


Rodrigo Agerri, Eneko Agirre, Itziar Aldabe, Nora Aranberri, Jose Maria Arriola, Aitziber Atutxa, Gorka Azkune, Arantza Casillas, Ainara Estarrona, Aritz Farwell, et al. 2021. European Language Equality – Deliverable D1.2 – Report on the state of the art in Language Technology and Language-centric AI, September 2021.

European Parliament. 2018. Language Equality in the Digital Age. European Parliament resolution of 11 September 2018 on Language Equality in the Digital Age (2018/2028(INI).

Frascati Manual. 2015. Guidelines for collecting and reporting data on research and experimental development.

Federico Gaspari, Annika Grützner-Zahn, Georg Rehm, Owen Gallagher, Maria Giagkou, Stelios Piperidis, and Andy Way. 2023. Digital Language Equality: Definition, Metric, Dashboard. In Georg Rehm and Andy Way, editors, *European Language Equality: A Strategic Agenda for Digital Language Equality*, Cognitive Technologies, pages 39–73. Springer, Cham, Switzerland.

Daniel Gildea, Min-Yen Kan, Nitin Madnani, Christoph Teichmann, and Martín Villalba. 2018. The ACL Anthology: Current state and future directions. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 23–28, Melbourne, Australia. Association for Computational Linguistics.

Lezan Hawizy, David M Jessop, Nico Adams, and Peter Murray-Rust. 2011. Chemicaltagger: A tool for semantic text-mining in chemistry. *Journal of cheminformatics*, 3:1–13.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Anne C Kroon, Toni van der Meer, and Rens Vliegenthart. 2022. Beyond counting words: Assessing performance of dictionaries, supervised machine learning, and embeddings in topic and frame classification. *Computational Communication Research*, 4(2):528–570.

Maite Melero, Pablo Peñarrubia, David Cabestany, Blanca C. Figueras, Mar Rodríguez, and Marta Villegas. 2022. European Language Equality – Deliverable D1.32 – Report on the Portuguese Language.

Ruslan Mitkov. 2022. *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Joerg Osterrieder. 2023. A primer on natural language processing for finance. *Available at SSRN 4317320*.

Daniel W Otter, Julian R Medina, and Jugal K Kalita. 2020. A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624.

Georg Rehm, editor. 2023. *European Language Grid: A Language Technology Platform for Multilingual Europe*. Cognitive Technologies. Springer, Cham, Switzerland.

Georg Rehm and Hans Uszkoreit, editors. 2012. *META-NET White Paper Series: Europe's Languages in the Digital Age*, 32 volumes on 31 European languages. Springer, Heidelberg etc.

Georg Rehm and Andy Way, editors. 2023. *European Language Equality: A Strategic Agenda for Digital Language Equality*. Cognitive Technologies. Springer, Cham, Switzerland.

Shaurya Rohatgi. 2022. ACL Anthology Corpus with Full Text. Github. Accessed: 2023-01-10.

Unesco. 1978. Recommendation concerning the international standardization of statistics on science and technology.

## A. List of Languages

1. Bulgarian
2. Catalan/Valencian
3. Croatian
4. Czech
5. Danish
6. Dutch
7. English
8. Estonian
9. Finnish
10. French
11. German
12. Greek
13. Hungarian
14. Irish
15. Italian
16. Latvian
17. Lithuanian
18. Maltese
19. Polish
20. Portuguese
21. Romanian/Moldavian/Moldovan
22. Slovak
23. Slovene
24. Spanish
25. Swedish
26. Basque
27. Bosnian
28. Faroese
29. Galician
30. Icelandic
31. Luxembourgish
32. Norwegian
33. Serbian
34. Tornedalian
35. Welsh
36. Karelian
37. Romani
38. Saami
39. Yiddish

## B. List of Domains based on FORD and ELE Classifications

| Class | Domain |
|---|---|
| Natural sciences | Mathematics |
| | Computer and information sciences |
| | Physics |
| | Chemistry |
| | Environmental sciences |
| | Biological sciences |
| Engineering and technology | Civil engineering |
| | Electrical engineering |
| | Electronic engineering |
| | Information engineering |
| | Mechanical engineering |
| | Chemical engineering |
| | Materials engineering |
| | Medical engineering |
| | Environmental engineering |
| | Environmental biotechnology |
| | Industrial biotechnology |
| | Nano-technology |
| Agricultural and veterinary sciences | Agriculture |
| | Forestry |
| | Fisheries |
| | Animal and dairy science |
| | Veterenary science |
| | Agricultural biotechnology |
| Social sciences | Psychology |
| | Cognitive sciences |
| | Economics |
| | Business |
| | Finance |
| | Tourism |
| | Education |
| | Sociology |
| | Law |
| | Political Science |
| | Government |
| | Geography |
| Humanities and the arts | History |
| | Archeology |
| | Anthropology |
| | Literature |
| | Philology |
| | Linguistics |
| | Philosophy |
| | Ethics |
| | Religion |
| | Arts |

Table 2: List of domains based on FORD and ELE classifications

## C. List of NLP Tasks

| Class | NLP Task |
|---|---|
| Text and speech processing | Optical character recognition |
| | Speech recognition |
| | Speech segmentation |
| | Text-to-speech |
| | Word segmentation (Tokenization) |
| Morphological analysis | Lemmatization |
| | Morphological segmentation |
| | Part-of-speech tagging |
| | Stemming |
| Syntactic analysis | Grammar induction |
| | Sentence breaking |
| | Parsing |
| Lexical semantics | Lexical semantics |
| | Distributional semantics |
| | Named entity recognition |
| | Sentiment analysis |
| | Terminology extraction |
| | Word-sense disambiguation |
| | Entity linking |
| | Multiword Expressions |
| Relational semantics | Relationship extraction |
| | Semantic parsing |
| | Semantic role labelling |
| Discourse | Coreference resolution |
| | Discourse analysis |
| | Implicit semantic role labelling |
| | Recognizing textual entailment |
| | Topic segmentation |
| | Argument mining |
| | Anaphora resolution |
| | Temporal processing |
| Higher-level NLP applications | Automatic summarization |
| | Grammatical error correction |
| | Machine translation |
| | Natural-language understanding |
| | Natural-language generation |
| | Book generation |
| | Document AI |
| | Dialogue management |
| | Question answering |
| | Text-to-image generation |
| | Text-to-scene generation |
| | Text-to-video |
| | Information retrieval |
| | Information extraction |
| | Multimodal systems |
| | Automated writing assistance |
| | Text simplification |
| | Author profiling |
| | Spam detection |
| | Virtual agents and chatbots |

Table 3: List of NLP tasks

## D. Number of Articles presenting Research about a certain Domain

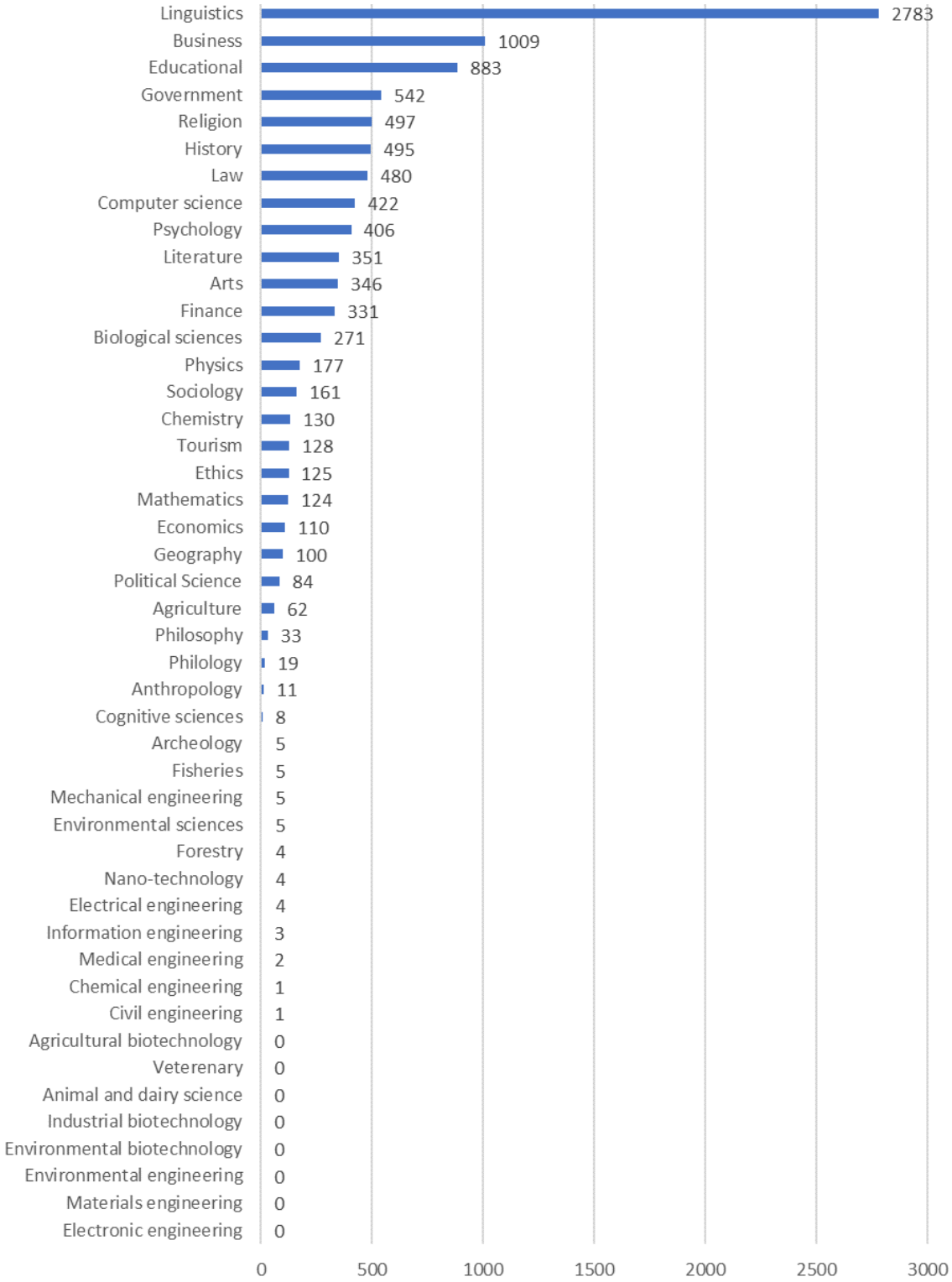


Figure 5: Number of articles presenting research about a certain domain

## E. Number of Articles presenting Research about a certain NLP Task

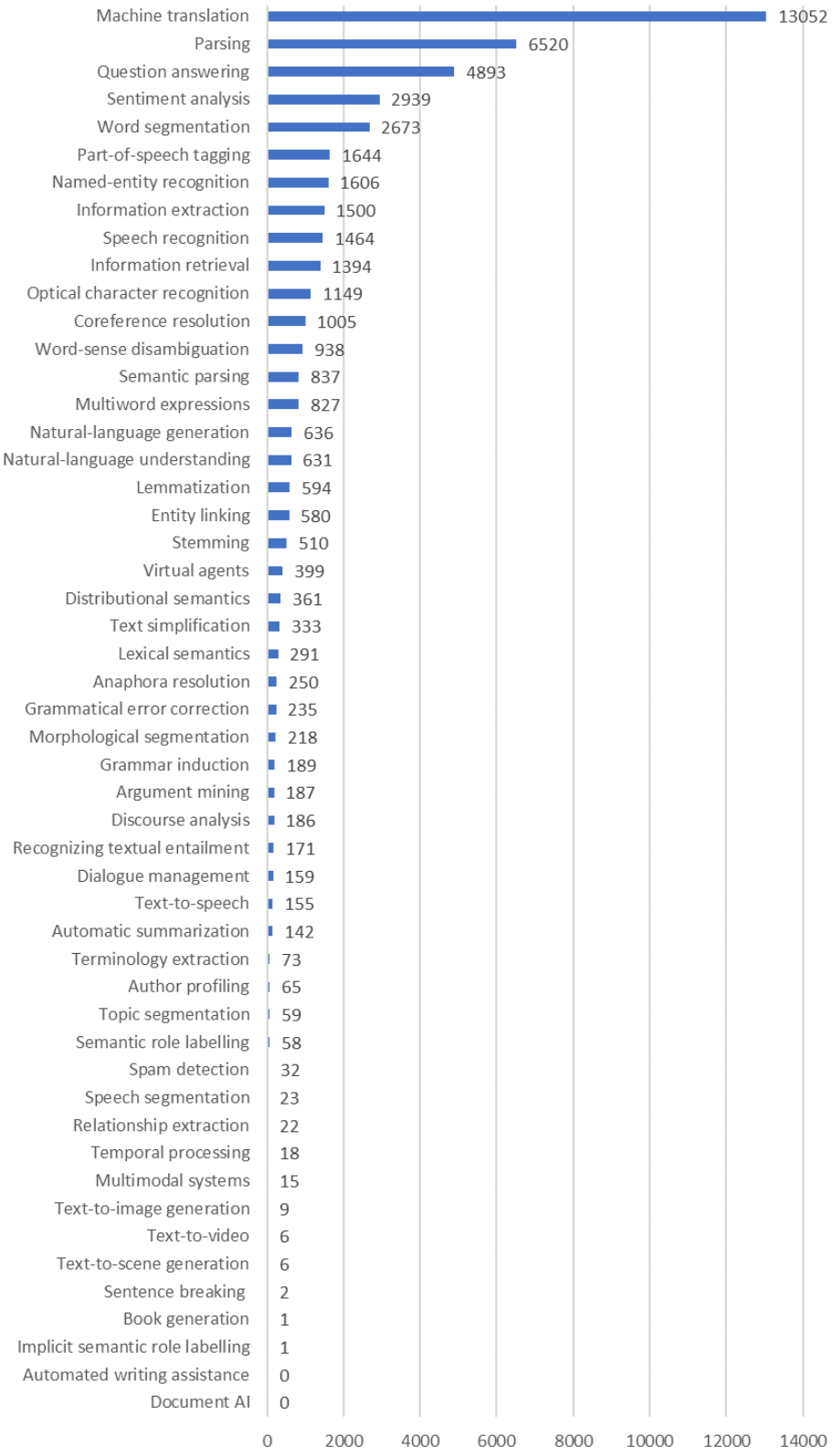


Figure 6: Number of articles presenting research about a certain NLP task

## F. Domains which are mostly associated with the Top 10 NLP Tasks

| NLP Task | Domains |
|---|---|
| Machine Translation | Arts, Biological Sciences, Business, Cognitive Sciences, Education, Ethics, Finance, Government, History, Law, Literature, Psychology, Religion, Sociology, and Tourism |
| Parsing | Arts, Biological Sciences, Business, Education, Finance, Government, History, Law, and Literature |
| Question Answering | Arts, Biological Sciences, Business, Education, Government, History, Law, and Religion |
| Sentiment Analysis | Business, Finance, Psychology, and Religion |
| Word Segmentation | Business, Education, Government, and Religion |
| Part-of-Speech tagging | Education |
| Named-entity recognition | Business |
| Information Extraction | Business and Government |
| Speech recognition | Business and Education |
| Information retrieval | Education |

Table 4: Domains which are mostly associated with the top 10 NLP tasks