EUROPEAN LANGUAGE EQUALITY 2

FSTP Project Report

EuLTDom2023 – European LT Domains 2023

Authors	Diego Alves, Marko Tadić
Organisation	University of Zagreb, Faculty of Humanities and Social Sciences
Dissemination level	Public
Date	31-03-2023

About this document

Project	European Language Equality 2 (ELE2)
Grant agreement no.	LC-01884166 – 101075356 ELE2
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-07-2022, 12 months
FSTP Project	EuLTDom2023 – European LT Domains 2023
Authors	Diego Alves, Marko Tadić
Organisation	University of Zagreb, Faculty of Humanities and Social Sciences
Type	Report
Number of pages	61
Status and version	Final
Dissemination level	Public
Date of delivery	31-03-2023
EC project officer	Susan Fraser
Contact	European Language Equality 2 (ELE2) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality 2 (ELE2) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu © 2023 ELE2 Consortium

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
5	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
6	European Federation of National Institutes for Language	EFNIL	LU
7	Réseau européen pour l'égalité des langues (European Language Equality Network)	ELEN	FR

Contents

List of Figures

List of Tables

List of Acronyms

ACL	Association for Computational Linguistics
AI	Artificial Intelligence
CL	Computational Linguistics
ELE	European Language Equality
ELG	European Language Grid (EU project, 2019-2022)
ELT	European Language Technology
FORD	Fields of Research and Development classification
LT	Language Technology/Technologies
NLP	Natural Language Processing
OECD	Organisation for Economic Co-operation and Development

Abstract

The EuLTDom2023 project reports on the current state-of-the-art of the usage of LT in different domains. The main purpose of this deliverable is to map the usage of NLP in various domains, to report findings regarding the fields that make regular use of these technologies, and to list domains that infrequently use LT or do not use it at all. For this aim, we analysed scientific papers published between 2010 and 2022 in the ACL Anthology, thus corresponding to scientific work done by the LT community. By applying a dictionary-based approach based on precise lists of terms related to languages, domains, and NLP tasks, we were able to present an overview of each of these dimensions, and to provide language-specific results mapping the usage of NLP tasks by the different domains. With the overall analysis, it is possible to identify the language, domains, and LT that have been the focus of the NLP research in the past years, and the language-specific results allow the clear identification of potential future developments to improve language equality at the level usage of LT.

1 Introduction

The fields of natural language processing (NLP) and Computational Linguistics (CL) encompass a large diversity of topics involving computational processing and the understanding of human languages. As described by (Agerri et al., 2021) while CL is more informed by linguists, NLP focuses more on the computation methods, Language Technology (LT) being a more neutral term. As it is not possible to define specific borders for each of them, and because there is a strong collaboration between the researchers from these fields, in this report, we treat them interchangeably.

Language Technology is part of everyone's life at different levels. Moreover, over the last several years, the field of natural language processing has been propelled forward by an explosion in the use of deep learning models (Otter et al., 2020) as these new architectures allowed the efficiency of language technologies to be highly improved.

However, even though the performance of LT has been deeply enhanced, many challenges still exist concerning the "Language equality in the digital age" resolution published by the European Parliament in 2018 which includes the support for the Human Language Project, formulated as the "establishment of a large-scale, long-term coordinated funding programme for research, development, and innovation in the field of Language Technology, at European, national and regional levels, tailored specifically to Europe's needs and demands as well as securing Europe's leadership in language-centric AI."

The main challenge regarding language equality is the threat of digital extinction of languages with smaller numbers of speakers (Rehm and Hegele, 2018). As the performance of the machine and deep learning methods rely on the existence of a large amount of data, these languages are commonly in a disadvantaged position. The situation of 39 European languages regarding the availability of LT was described in the reports[1] of the first European Language Equality (ELE) project.

Besides understanding the specific needs of each language, continuous mapping of the general NLP landscape is certainly one of the crucial steps for achieving digital language equality in the near future. While initiatives such as the European Language Grid (ELG) promote the deployment of state-of-the-art LT across languages (Rehm, 2023), a better comprehension of how different domains use the available NLP tools is crucial for identifying areas for improvement and opportunities for new research.

Handbooks describing LT usually focus on the technologies themselves, not on their specific applications in different domains. It is the case of the " Handbook of Natural Language

[1] Available at: https://european-language-equality.eu/deliverables/

Processing" (Indurkhya and Damerau, 2010) and "The Oxford Handbook of Computational Linguistics" (Mitkov, 2022). Moreover, in terms of surveys, in most cases, they focus on a general analysis of the state-of-the-art concerning algorithms and performances (e.g., Otter et al., 2020), with no information on how the technologies are deployed by different domains.

On the other hand, it is possible to identify specific literature concerning the domains. For example, the article "A Primer on Natural Language Processing for Finance" (Osterrieder, 2023) presents a complete overview of the usages of LT in Finance. Furthermore, there are numerous works presenting NLP tools especially conceived for applications in certain domains, and other research papers from different fields where LT is part of the methodological process.

As of today, there is no broad comparative analysis of the deployment of LT in different fields, thus, the aim of this report is to map this usage in various domains. The idea is to answer the question: In which domains is NLP used a lot and in which, rarely or not at all?

Our purpose is to create a snapshot showing how LT is used in different fields and to detail the findings regarding the fields that make regular use of it for each one of the 39 European languages which were objects of analysis in the ELE project. In addition, also enlist the domains that infrequently use NLP or do not use it at all. The fundamental fragmentation that exists among the LT community in Europe favours some domains over others. Thus, a better understanding of the current efforts in terms of highly sought-after LT for specific domains per language will enable the required action to support the underdeveloped or non-developed ones. Providing grounds for more efficient redirection of efforts to cover LT for all domains could directly lead to the widening of the European LT communities, and to the creation of newer opportunities for underdeveloped domains in different languages and countries.

For this aim, we propose to examine how the LT research community addresses its research in different domains, favouring some of them and neglecting others. The idea is to analyse the presence of the selected domains in LT research papers available in the Association for Computational Linguistics (ACL) Anthology. As our focus is on the state-of-the-art in terms of LT, we decided to consider the research works published from 2010 to the end of 2022.

As the focal point of this report is on papers publicly available, it excludes research works done by private companies which do not publish their results due to internal privacy policies.

The report is composed as follows: Section 2 describes the data provided by the ACL Anthology. In Section 3, we detail the methodology for the information extraction and treatment, then, in Section 4, we present an overview of the languages, domains, and NLP tasks in the ACL Anthology, followed by a detailed analysis of the usage of domains and LT per language. Finally, Section 5 presents the conclusions with a focus on the opportunities regarding the underdeveloped domains and the usage of LT.

2 ACL Anthology

The Association for Computational Linguistics (ACL) is the main international scientific and professional society for researchers working on language technologies. It was founded in 1962 and its activities include holding an annual meeting each summer and sponsoring the journal Computational Linguistics, published by MIT Press, as well as compiling published NLP articles from ACL and non-ACL events (i. e., ACL Anthology[2]) (ACL).

The ACL Anthology is a key Open Access archive with Open Source components for researchers in the NLP community. It is the main source of computational linguistics and natural language processing scientific literature, currently maintained exclusively by community

[2] Available at: https://aclanthology.org/

volunteers. It offers both text and faceted search of the indexed papers and author-specific pages, and it allows open access to the proceedings of all ACL-sponsored conferences and journal articles, also hosting third-party computational linguistics literature from sister organizations and their national venues. (Gildea et al., 2018). The coverage of this Anthology is presented below:

- ACL events: AACL, ACL, ANLP, CL, CoNLL, EACL, EMNLP, Findings, IWSLT, NAACL, SemEval,*SEM, TACL, WMT, WS, SIGs[3]
- Non-ACL events: ALTA, AMTA, CCL, COLING, EAMT, HLT, IJCLCLP, IJCNLP, JEP/TALN/RECITAL, KONVENS, LILT, LREC, MTSummit, MUC, NEJLT, PACLIC, RANLP, ROCLING, TAL, TINLAP, TIPSTER

For this report, we used the ACL Anthology Corpus repository which provides PDF files, full-text, references, and other details extracted by GROBID (GRO, 2008–2023) from the PDF files available in the ACL Anthology (Rohatgi, 2022). It contains data from 80,013 ACL articles and posters from 1957 to October 2022.

As previously mentioned, we are focusing on more recent research works to examine how different languages, domains, and NLP tasks are distributed throughout the papers. Thus, we created a sub-set of the ACL Anthology which is composed of texts published between 2010 and October 2022, a total of 49,466 articles. The distribution of the number of articles per year is presented in Figure 1.

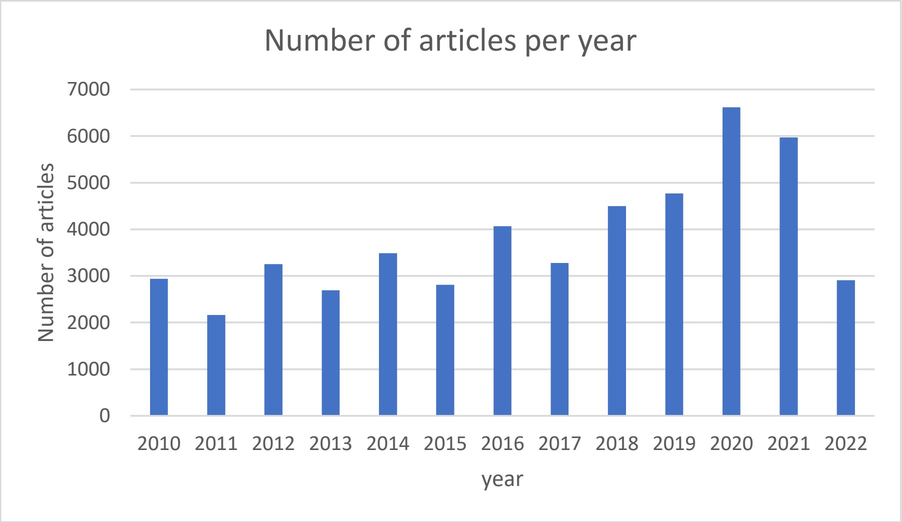

Figure 1: Number of articles from the ACL Anthology per year between 2010 and 2022.

The number of articles per year varies from 2,000 to 5,000, with the exception of 2020 and 2021, with more publications compared to the other years. Not all conferences happen every year, it is the case of LREC and COLING which are biannual (i. e., every even year).

[3] SIG stands for Special Interest Group, a collection of proceedings for specific conferences and workshops related to some NLP tasks.

Although being the main reference regarding LT publications, the ACL Anthology does not contain all the research work developed by the NLP community worldwide, and the articles are mainly in English. Thus, it does not encompass all the research that is developed locally targeting the specific needs of certain communities. However, concerning the scope of this project, it provides considerably valuable information regarding the LT state-of-the-art for the analysis of its usage by different domains.

3 Methodology

The methodology adopted for this work is schematised in Figure 2. To understand the usage of LT by different domains in different languages, we decided to use a dictionary approach. The idea is to count the number of research works in our subset of the ACL Anthology Corpus (i. e., articles from 2010 to 2023) that mention the defined terms regarding languages, domains, and NLP tasks.

The first task concerns an overall analysis of each one of these dimensions separately, allowing us to identify the languages, domains, and LT tasks that are favoured in NLP articles, and to point out the ones underrepresented.

The second task is conducted separately for each language. In this step, we count the number of articles that mention the domain/NLP task pair, allowing us to clearly identify how different domains are using each specific LT.

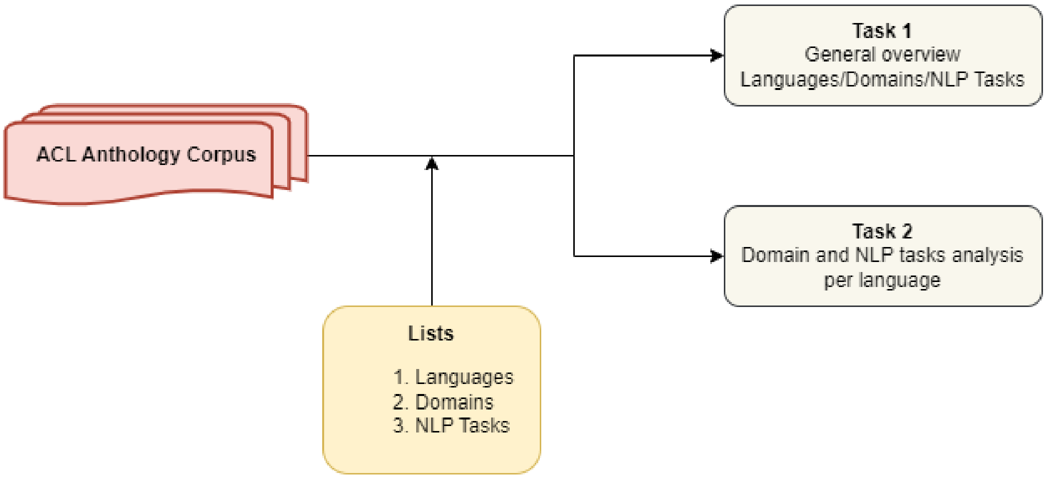

Figure 2: Methodological Schema.

Before analysing the ACL Anthology Corpus texts, we defined the lists of languages, domains, and NLP tasks according to precise criteria as described in the following subsections.

3.1 Languages

Our language set is composed of European languages which have an ELE Language Report update as described in the Introduction section.

In total, 39 languages were selected: Bulgarian, Catalan/Valencian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian/Moldavian/Moldovan, Slovak, Slovene, Spanish, Swedish, Basque, Bosnian, Faroese, Galician, Icelandic, Luxembourgish, Norwegian, Serbian, Tornedalian, Welsh, Karelian, Romani, Saami, Yiddish.

For the languages which have more than one denomination (i. e., Catalan/Valencian and Romanian/Moldavian/Moldovan), while searching for the number of mentions in each paper, we considered all the possible terms referring to them.

3.2 Domains

To determine the list of relevant domains, we decided to use, as our main reference, the "Fields of Research and Development classification" (FORD), which is recommended by the Organisation for Economic Co-operation and Development (OECD) and which was the base of the "Frascati Manual 2015 – Guidelines for Collecting and Reporting Data on Research and Experimental Development" (Manual et al., 2015).

The FORD classification was developed for Research and Development measurement purposes and follows a content approach. The domains are grouped together to form broad (one-digit) and narrower (two-digit) fields of the classification. In addition, this approach for defining domains is closely related to and consistent with UNESCO's "Recommendation concerning the International Standardisation of Statistics on Science and Technology" (Unesco, 1978).

When comparing the FORD classification with the list of domains selected for the composition of ELE language reports (e. g., D1.32 – Report on the Spanish Language (Melero et al., 2022)), it is possible to find a high level of correspondence. The ELE list is shorter and possesses some general terms such as "Technology", "Science", and "Innovation".

Thus, based on the FORD classification, we defined the list of relevant domains for this study by:

- Completing the list with ELE fields that are not present in the FORD one, excluding the generic terms mentioned above
- Removing the FORD elements that correspond to the category "Other" such as: "Other natural sciences"
- Excluding Health and Media domains as they were objects of other specific projects in this call
- Replacing "Economic geography" and "Social Geography" simply with "Geography" as, in preliminary tests, these more specific terms were not identified in ACL texts.

Table 1 presents the established classification of domains. In total 46 fields were selected, and divided into 5 broader classes.

3.3 NLP Tasks

For the composition of the list of LT, we combined the information provided by "The Oxford Handbook of Computational Linguistics" (Mitkov, 2022) with additional tasks identified in the article "Natural Language Processing" at the Wikipedia website[4] (Wikipedia).

While the Oxford Handbook divides LT into two main categories (i. e., tasks and applications), Wikipedia has a more detailed classification which we decided to keep in this study. Moreover, we decided to add two tasks that are not described in the previously mentioned references but that are mentioned on the IBM website[5]: "Spam detection" and "Virtual agents and chatbots" (IBM).

Thus, our final list of LT is composed of 51 tasks divided into 7 categories and is presented in Table 2.

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

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

Domains	
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 1: List of domains based on FORD and ELE classifications.

Domains	
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 2: List of NLP tasks.

3.4 Text Processing

Once the lists of languages, domains, and NLP tasks were defined, we created the strategy for extracting information from the ACL Anthology corpus.

In the first task, the idea is to analyse each dimension separately. For that aim, we examined if each term in the above-mentioned lists was mentioned in each text of the collection, identifying, in the end, the total number of articles dealing with each element. Texts and query terms are in lowercase.

We used Python Regular expression operations library[6] which enables users to find all occurrences of a certain term in a determined text.

One of the problems concerning this method is the fact that a certain term may be mentioned in the article even if the text is not exactly focusing on it specifically. Therefore, we decided to define a threshold to reduce this bias. Thus, we consider that an article deals with each language/domain/NLP task if the term is mentioned at least twice in the text.

For each text available in the ACL Anthology Corpus, we consider its full text (i. e., from abstract to conclusion). The author's information and references were excluded. Unfortunately, not all conferences require authors to define keywords, this information would have been relevant to the type of study presented in this report.

For the second task, we examined languages separately. Thus, first, we identified if the text mentioned the language, then, we checked if the article mentioned each domain/NLP task pair. With the obtained information, we generated heat-maps with the seaborn statistical data visualization Python library[7] that allows easy identification of how domains use different types of LT.

Regarding the domains and LT tasks, the query for the mentions in the texts was conducted with the terms listed in Tables 1 and 2 but also with possible synonyms or other orthographic forms. The complete lists with the searched terms are presented in the digital material accompanying this report and accessible at URL[8].

Besides that, some terms in the list of domains may be used in different contexts not specifically referring to the domain. It is the case for example of "literature" and "history". In these cases, for the domain to be considered, besides the noun, also the respective adjective must be mentioned in the text for the article to be counted (e. g., literature and literary; history and historical).

A specific treatment had to be implemented for the domain "Arts" (or "Art"). As many papers contain the term "state-of-the-art" or its variations, we established, in our script, a way to verify the context of the match to guarantee that the count does not consider this phrase.

The dictionary approach described here relies on the count of the occurrences of specific terms in the texts. This approach presents some weaknesses when compared to methods for topic classification based on Supervised Machine Learning methods, and embeddings Kroon et al. (2022). Considering the restrictions of this project in terms of time and resources, the selected approach was the optimal choice to provide the desired snapshot regarding the usage of LT by different domains for the different languages in our language set.

4 Results

In this section, we present the results regarding each one of the tasks identified in Figure 2, starting with an overall analysis concerning each dimension of this study (i. e., languages,

[6] https://docs.python.org/3/library/re.html
[7] https://seaborn.pydata.org/
[8] https://github.com/dfvalio/EuLTDom2023

domains, and NLP tasks), followed by a specific study of the usage of LT by different domains per language.

4.1 Overall Analysis

4.1.1 Languages

The results concerning the number of ACL articles mentioning at least twice each language in our set are presented in Figure 3. The only language which never appears in our data is Tornedalian.

Of the 49,466 texts in the ACL Anthology Corpus (from 2010 to 2022), 45,737 (92,46%) mention at least one of the languages from our language set twice which shows a predominance of European languages in the NLP studies. However, they are not distributed homogeneously.

As expected, the most mentioned language is by far English (i. e., more than 20,000 articles), followed by German, French, and Spanish (all with more than 3,000 articles). These results are coherent with similar studies such as "The State and Fate of Linguistic Diversity and Inclusion in the NLP World'' (Joshi et al., 2020) which presents an analysis in terms of entropy of the disparity between languages, especially in terms of their LT using an older version of the ACL Anthology.

Italian, Czech, and Portuguese present intermediate results, from 1,000 to 1,500 articles, and the vast majority of languages are mentioned in a number of articles comprised between 100 to 1,000.

And the languages with the smallest representation in the NLP research works published between 2010 and 2022 are Galician, Welsh, Maltese, Bosnian, Faroese, Saami, Karelian, Yiddish, Luxembourgish, and Tornedalian.

These are general numbers concerning all the conferences in the ACL Anthology database. Joshi et al. (2020) showed that some events such as LREC tend to have more linguistic diversity in comparison to 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.

4.1.2 Domains

The overview of the number of articles mentioning at least twice each domain is presented in Figure 4. From the ensemble of articles in ACL published between 2010 and 2022, only 6,179 (12,49%) clearly mention the selected terms in our domain list. This may be due to the fact that the focus of many articles is on the development of the LT itself, not on its applications and to the exclusion of health and media domains from our scope. Additionally, some papers may concern the domains considered here, although not referring to them directly. These specific cases cannot be examined with our dictionary-based approach.

Linguistics is the most cited domain which is an expected result as our data concerns work published in Computational Linguistics conferences. The top ten most mentioned terms are from the Social Sciences and Humanities and the arts categories (varying from 2,783 to 351 articles). The first domain from a different category to appear in the figure is Biological sciences and is followed by other Natural Sciences domains such as Physics, Chemistry, and Mathematics. Engineering and technology is the category with the least number of articles.

A clearer view of the distribution of the domain categories is displayed in Figure 5. Together, the Social sciences and Humanities and arts correspond to 89,95% of the mentions. Thus, although presenting relevant work regarding new technologies in the computational domain, most papers seem to focus on the application for the social and humanities disciplines where language is either a direct object of their research or their object of research is predominantly mediated through language.

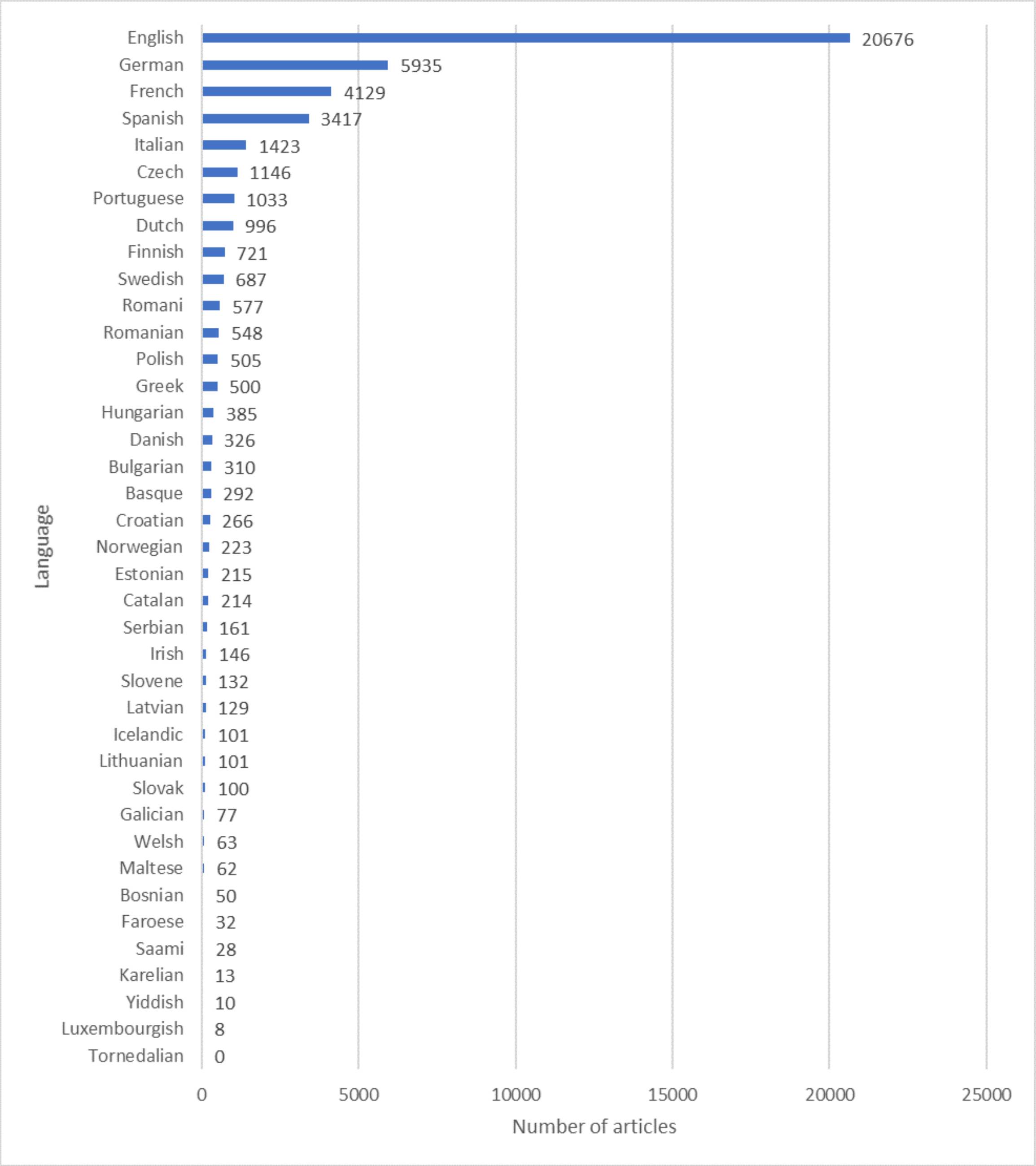

Figure 3: Number of articles presenting research about a certain language.

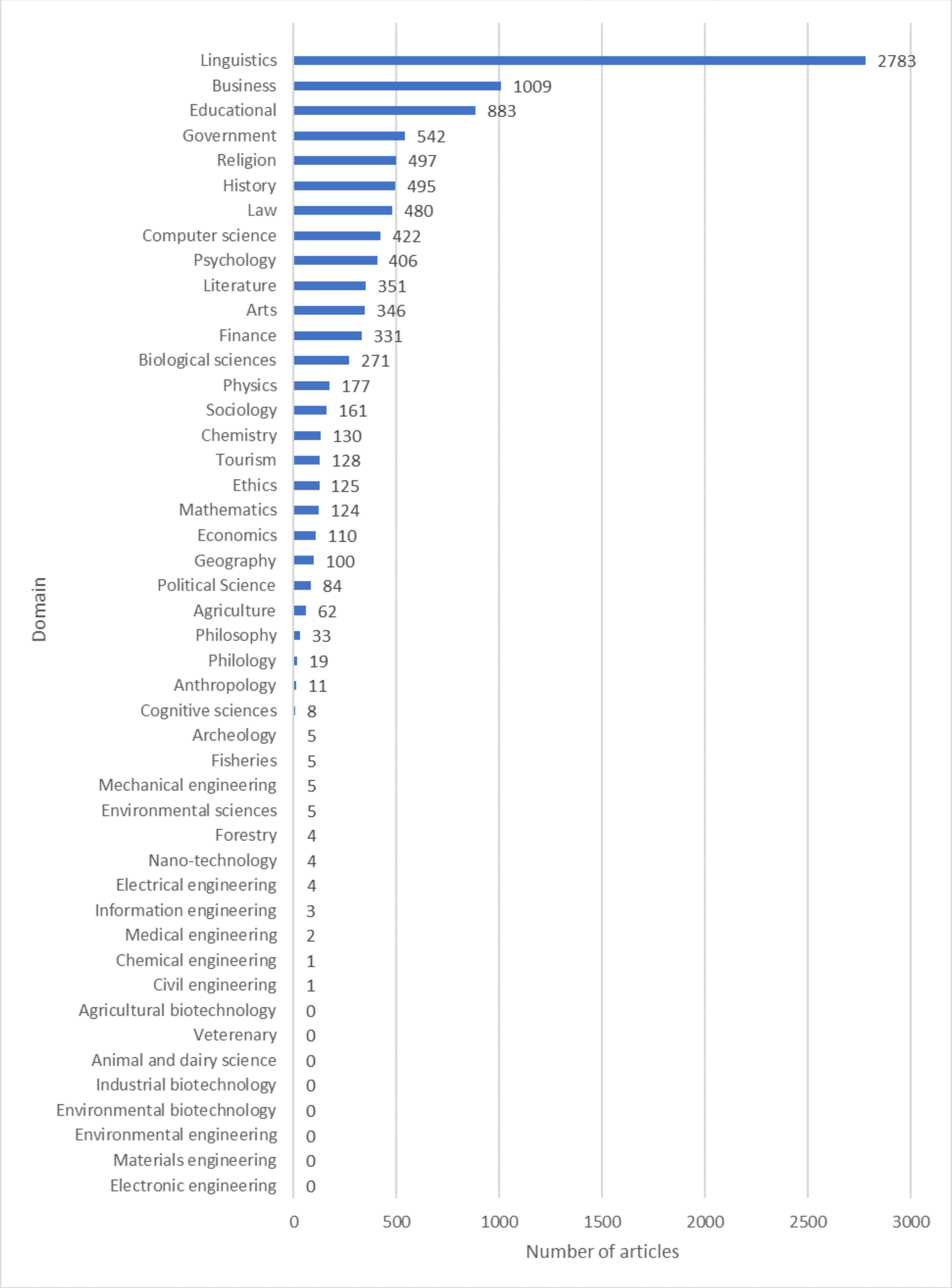

Figure 4: Number of articles presenting research about a certain domain.

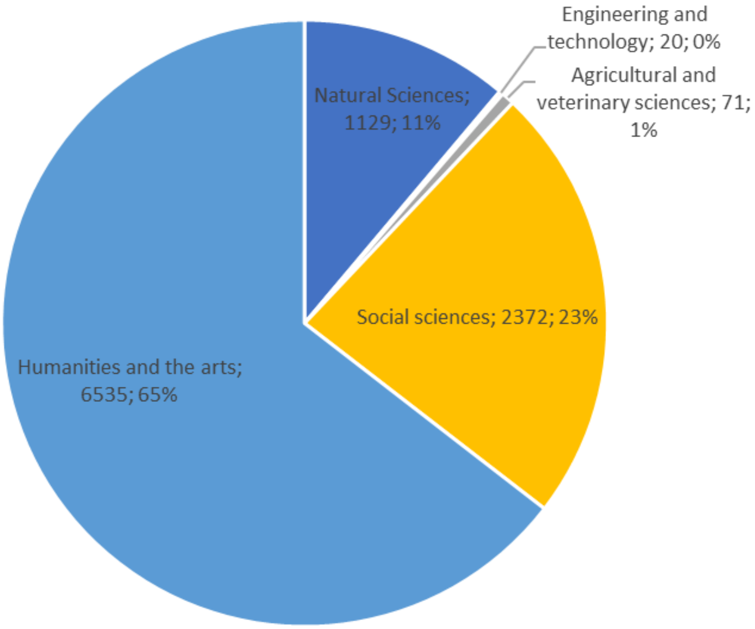

Figure 5: Number of articles presenting research about a certain domain category.

4.1.3 NLP Tasks

The analysis of the distribution of the NLP tasks in the ACL articles is presented in Figure 6. In total, 32,154 (65,00%) articles mention one of the tasks at least twice. One of the reasons for this percentage is the fact that some of the research work may involve other LT which were not considered in our NLP tasks list, or some authors may use other terms to refer to them.

From Figure 6, it is possible to notice that the LT research community has put a lot of effort into the Machine translation task. Its size is double the one of the second most cited tasks (Parsing). Another well-positioned Higher-level NLP application task is Question answering.

Furthermore, we can observe that tasks such as Parsing, Word Segmentation, Part-of-speech tagging, and Named-entity recognition are positioned in the top 10 of the most cited ones. This can be due to the fact that these operations are also part of more complex LT, being integrated into pipelines.

Of the 51 NLP tasks, 39 (76,47%) are mentioned in less than 1,000 articles, thus, presenting a lot of potential for further developments, for example, deployment of the existing architectures for languages other than English.

In Figure 7, we present the distribution of the categories regarding the NLP tasks. Almost half of the total number corresponds to Higher-level NLP applications (and half of it, to Machine Translation). The other three main categories are Lexical semantics, Syntactic Analysis, and Text and speech processing.

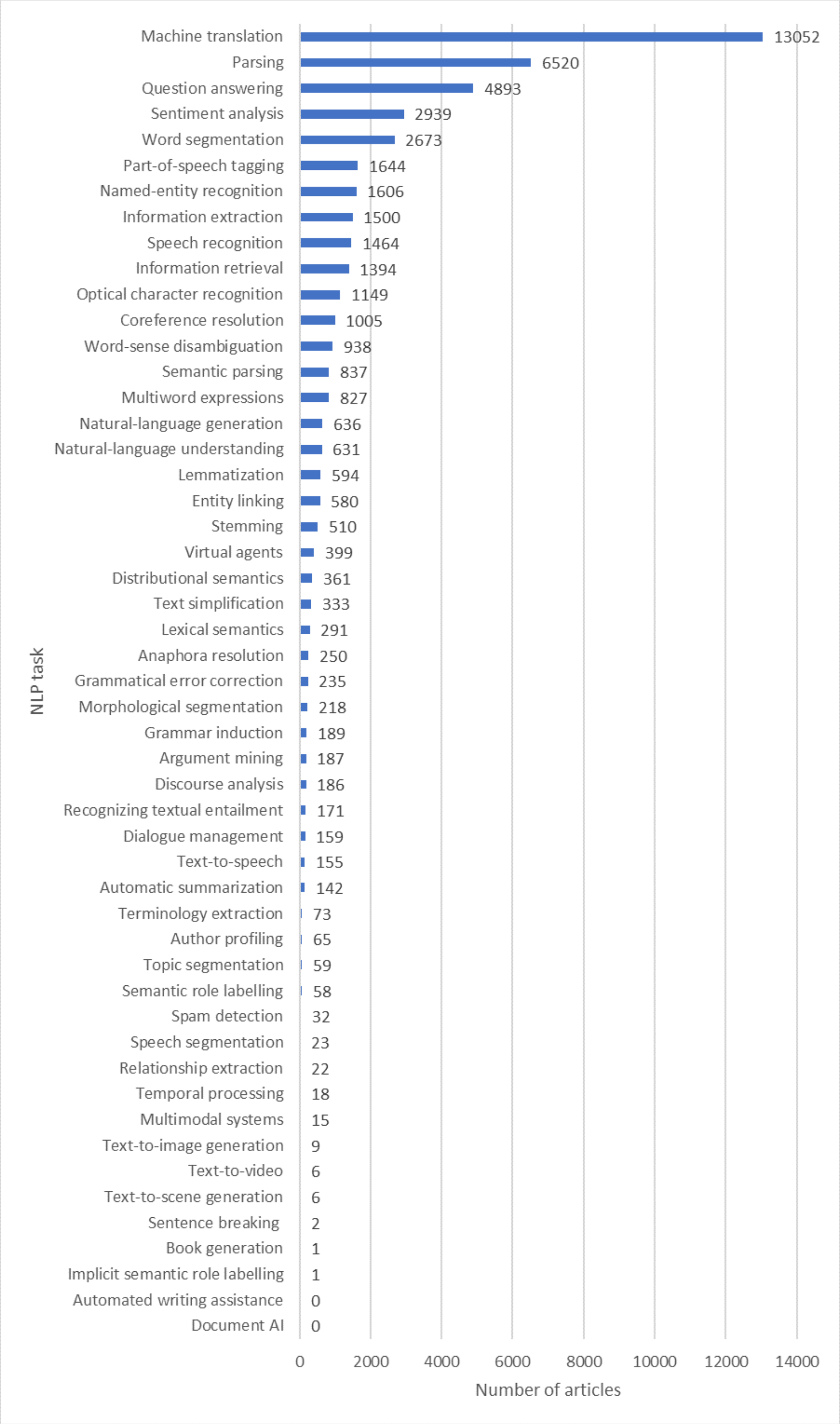

Figure 6: Number of articles presenting research about a certain NLP task.

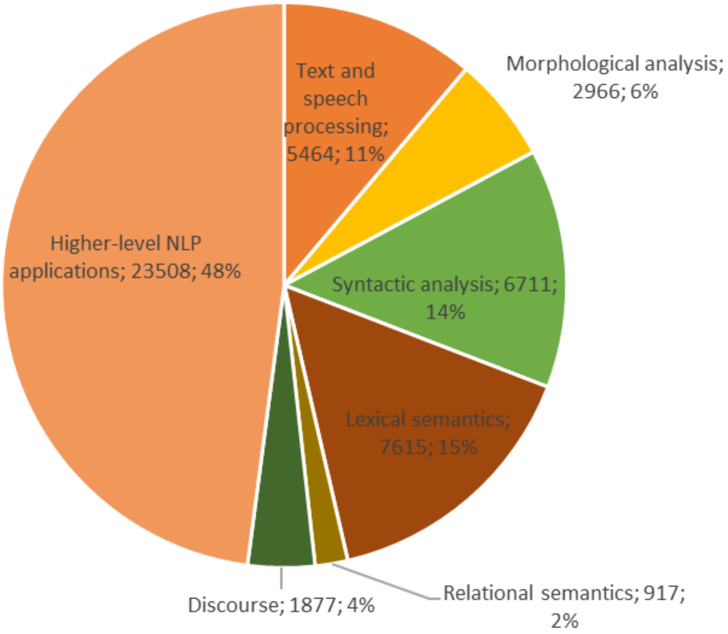

Figure 7: Number of articles presenting research about a certain NLP task category.

4.2 LT usage per domain

In this section, we present the analysis concerning the usage of LT by different domains per language (i. e., the number of articles where both domain and NLP task are mentioned at least twice each) and the results are commented in the “Discussion” subsection.

The extracted data per language (CSV files) are also available in the digital material accompanying this report[9]. Moreover, we also provide the svg files concerning the heat-maps as presented in the report.

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

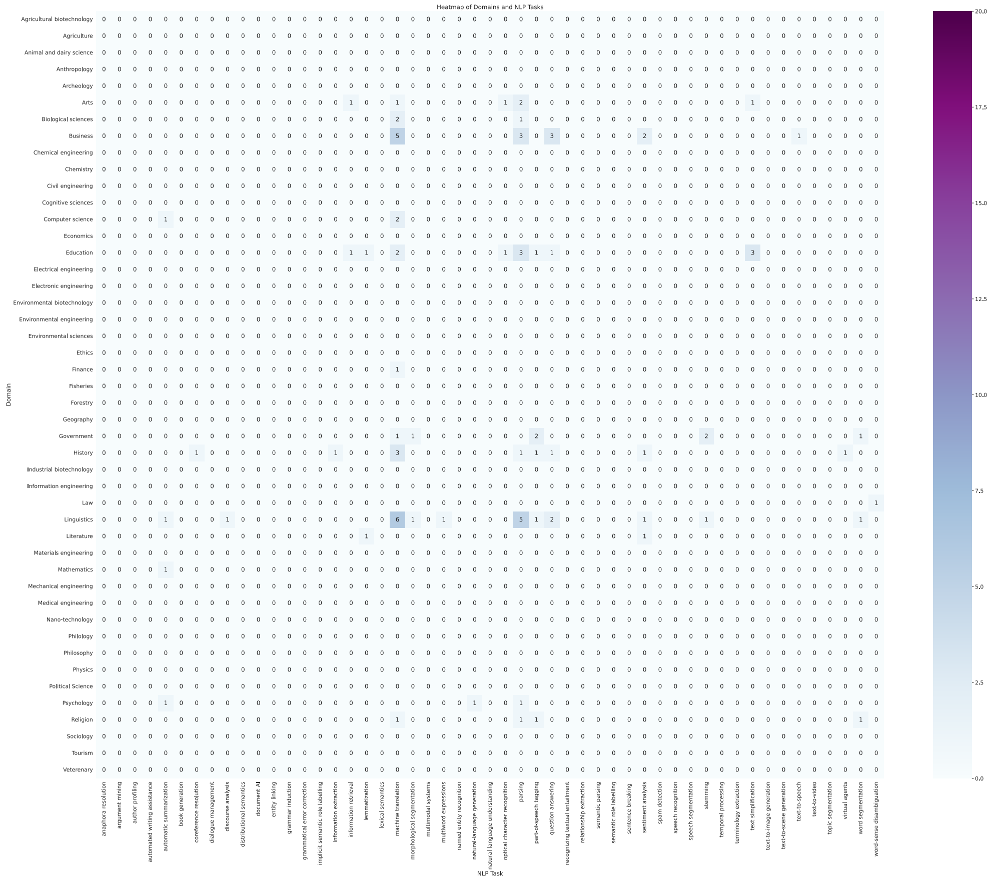

Figure 8: Number of articles per Domain and LT for Basque.

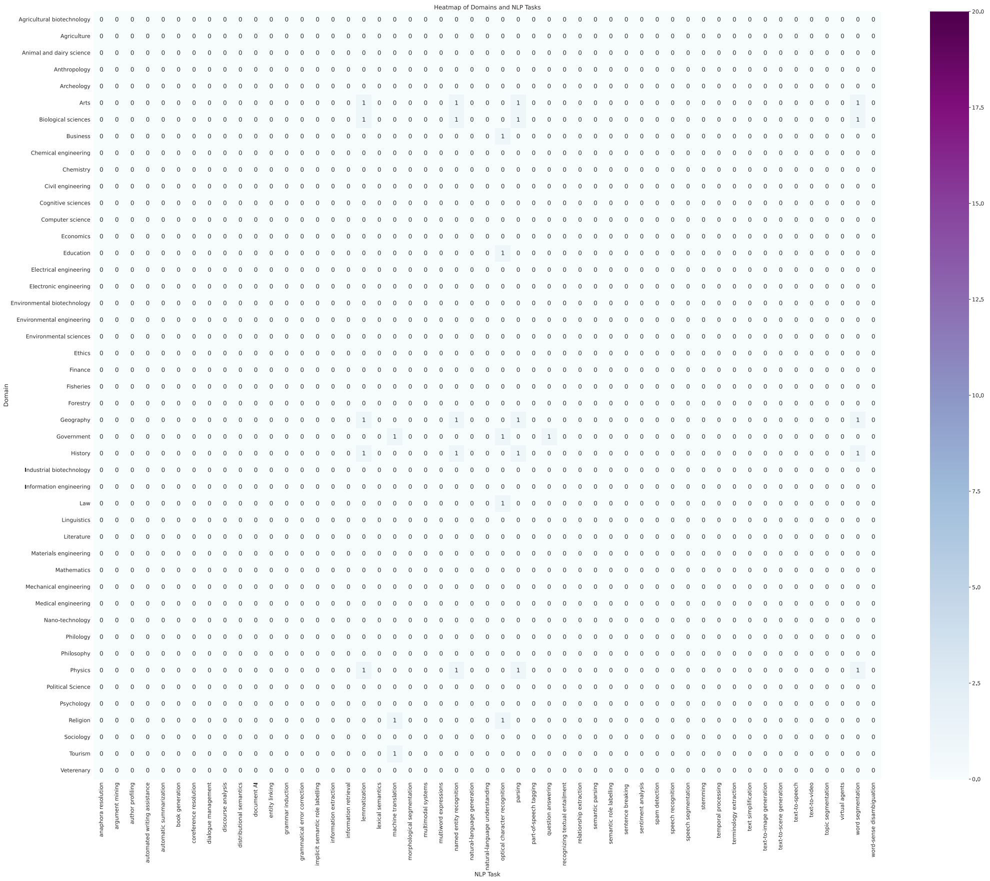

Figure 9: Number of articles per Domain and LT for Bosnian.

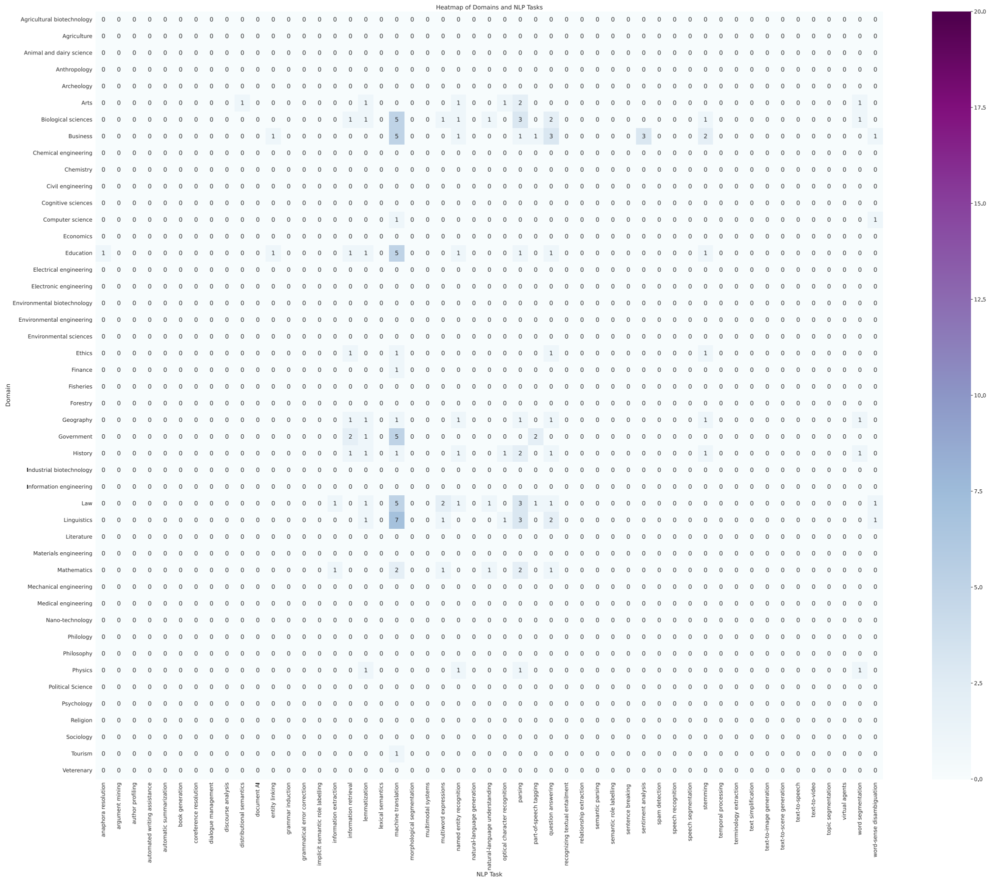

Figure 10: Number of articles per Domain and LT for Bulgarian.

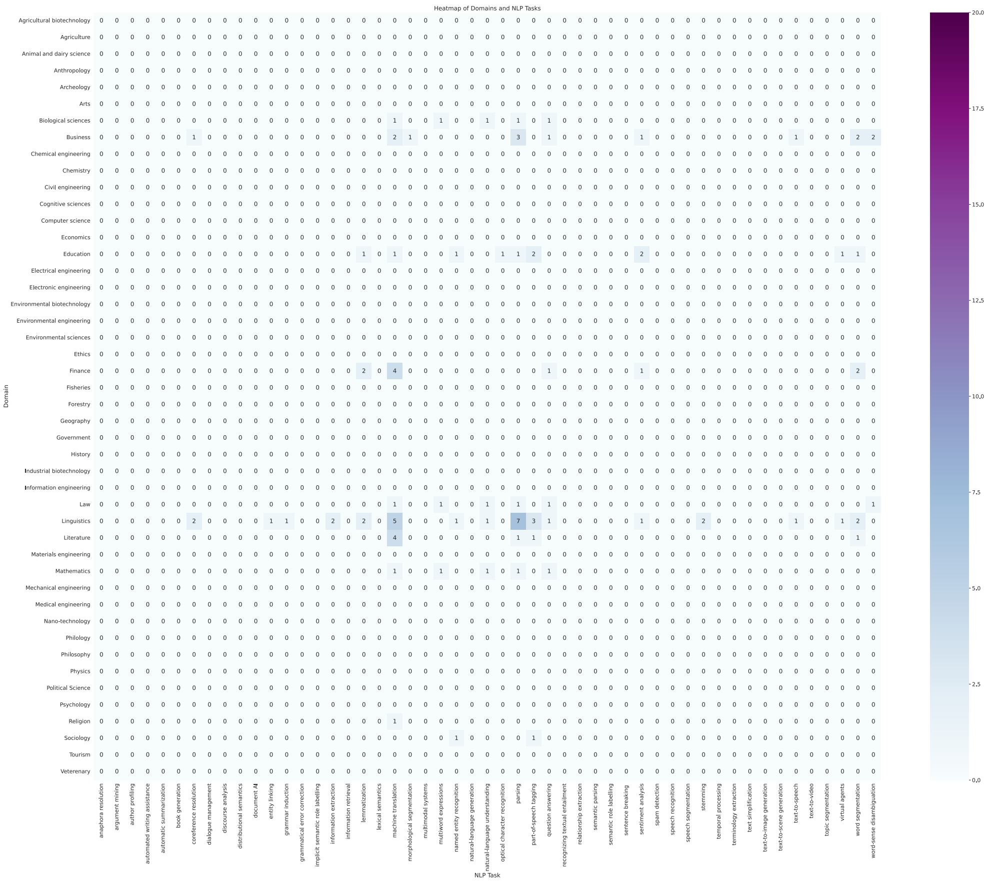

Figure 11: Number of articles per Domain and LT for Catalan.

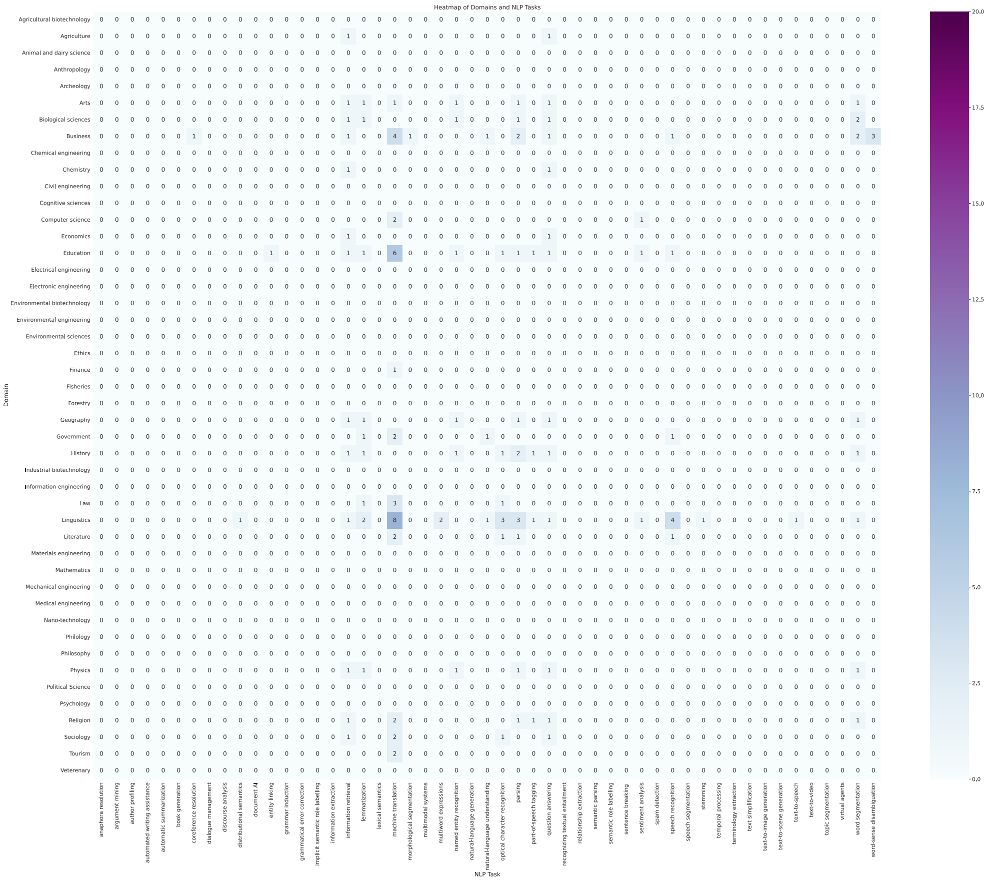

Figure 12: Number of articles per Domain and LT for Croatian.

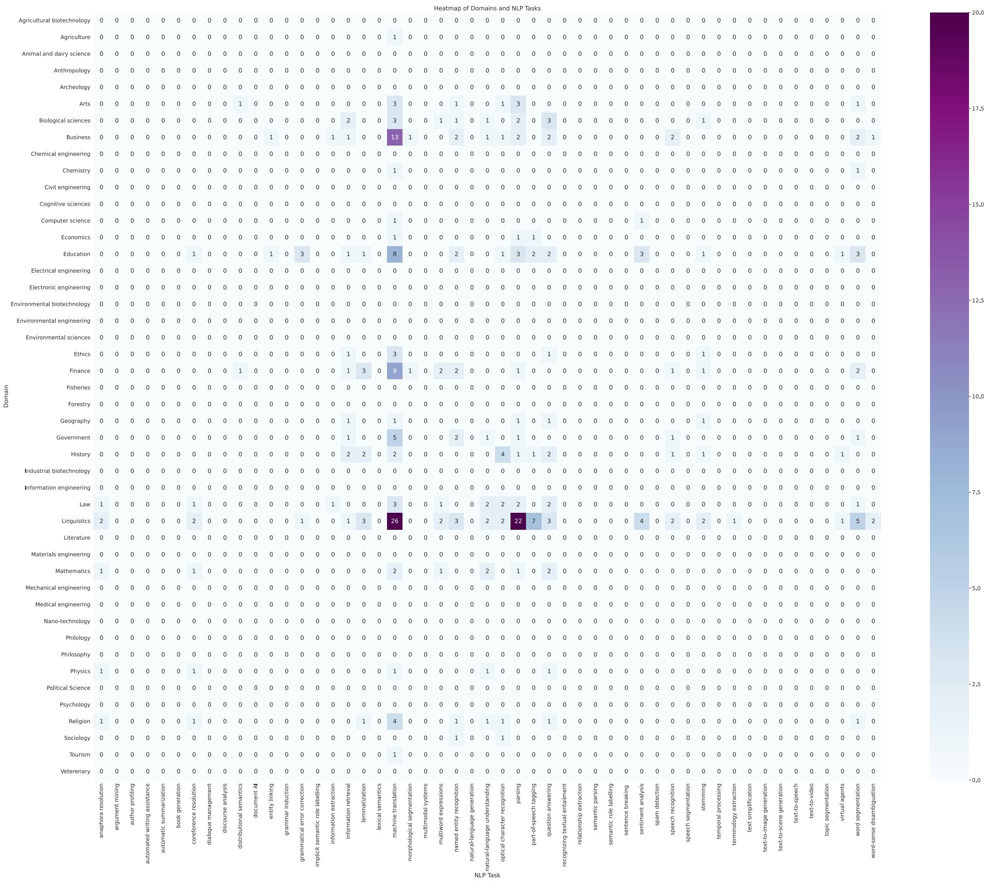

Figure 13: Number of articles per Domain and LT for Czech.

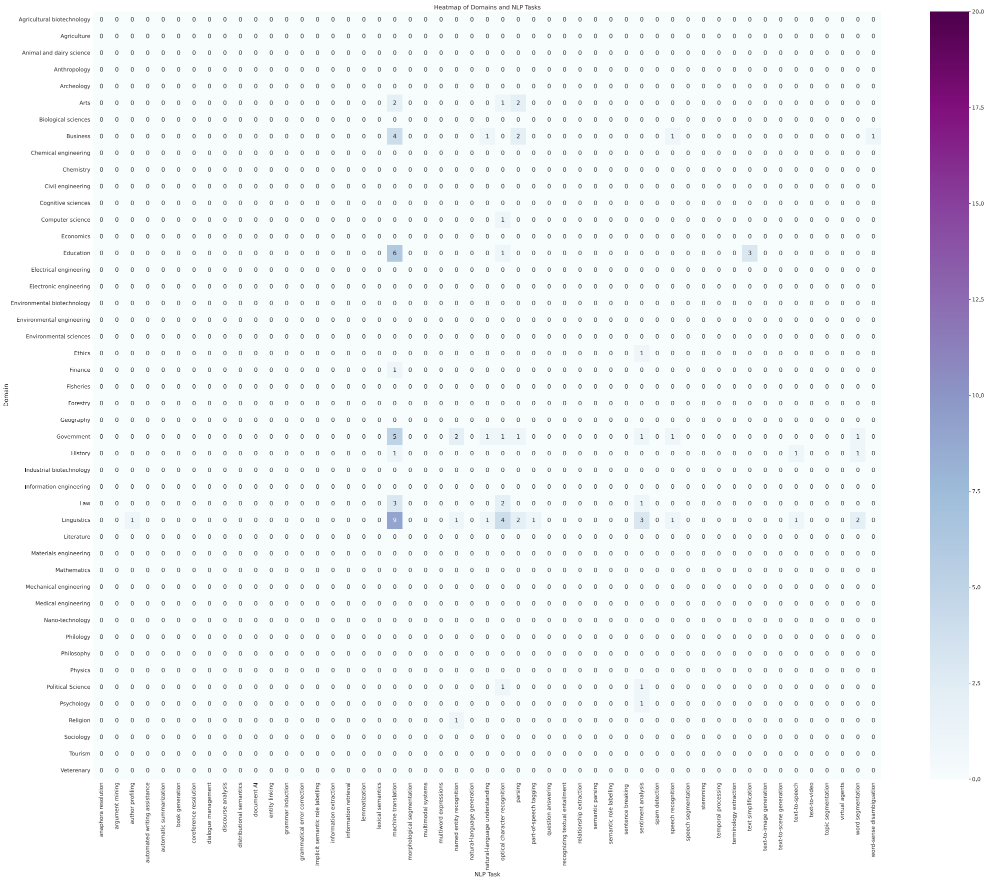

Figure 14: Number of articles per Domain and LT for Danish.

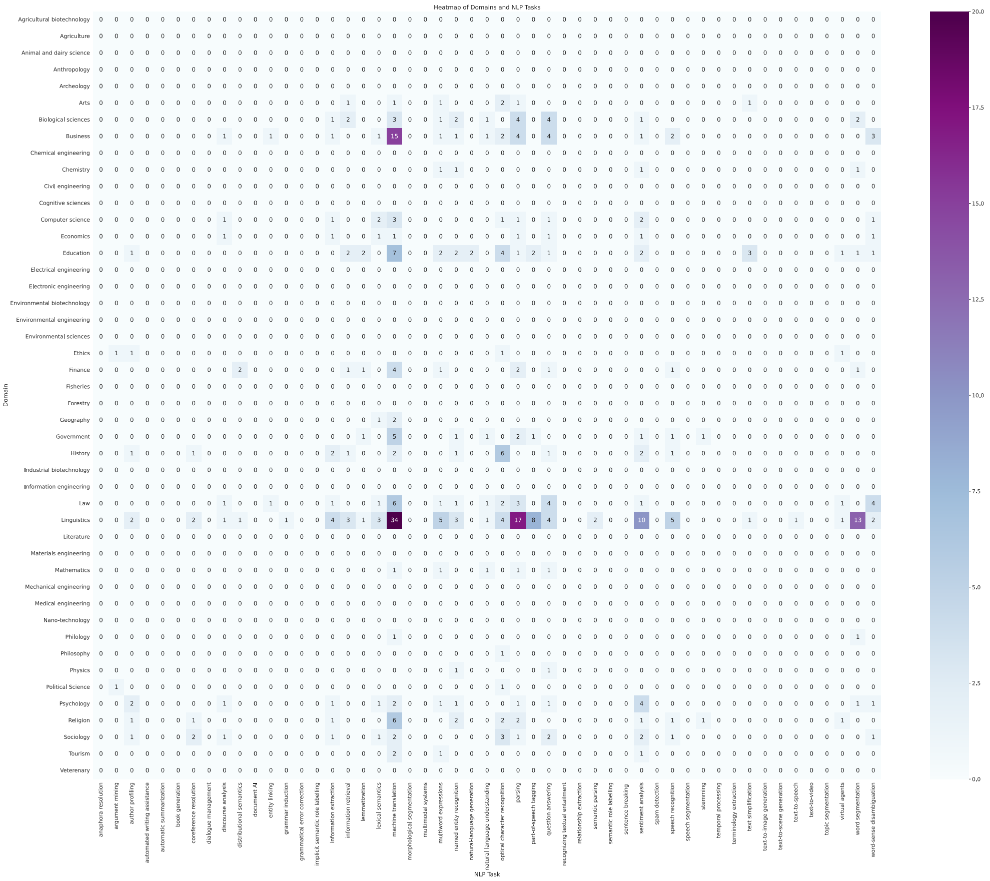

Figure 15: Number of articles per Domain and LT for Dutch.

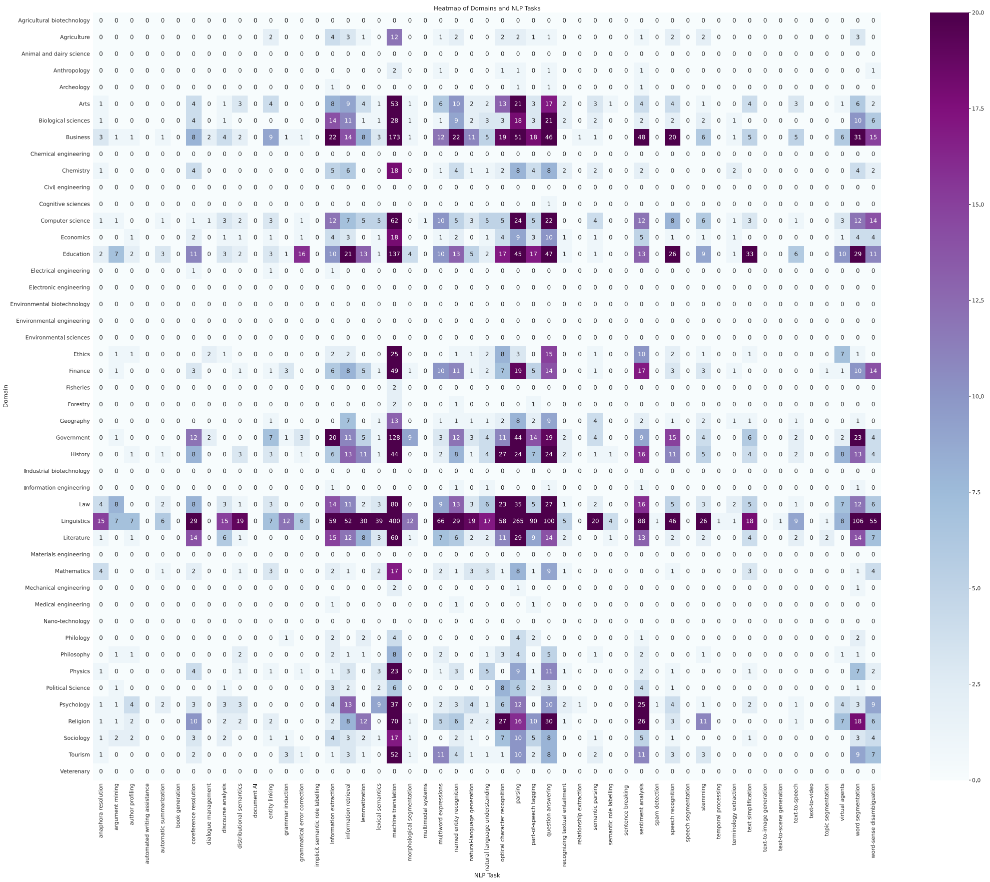

Figure 16: Number of articles per Domain and LT for English.

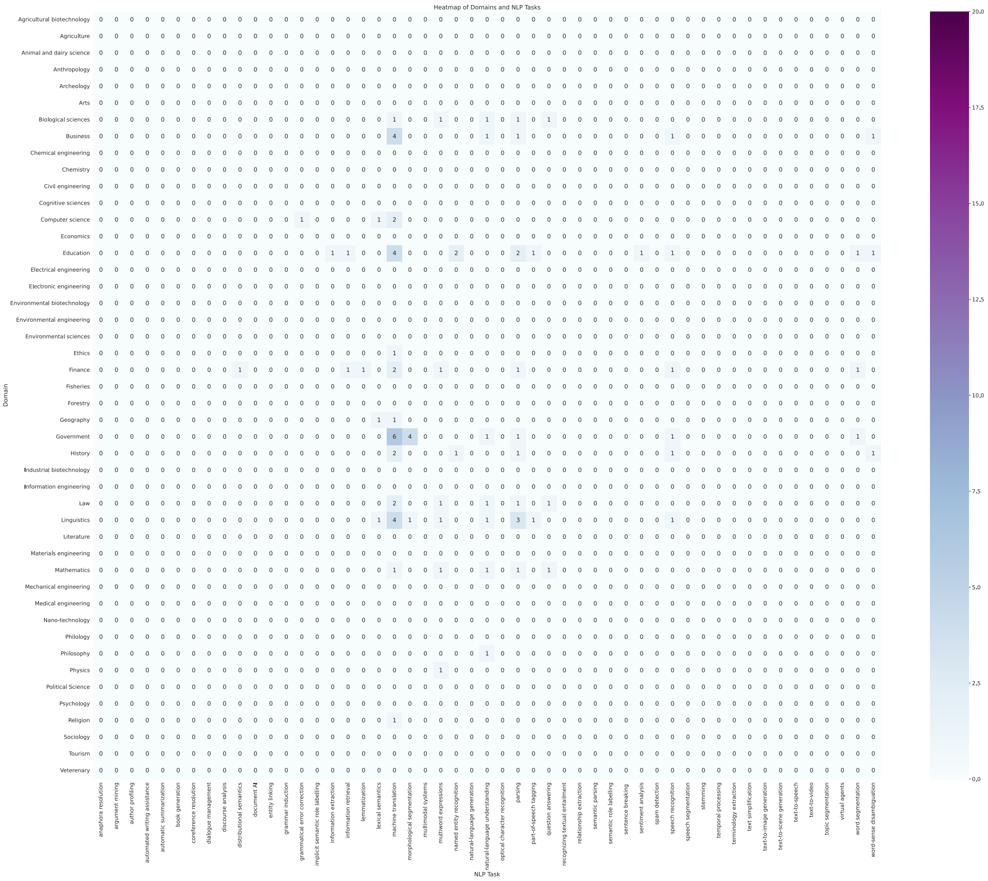

Figure 17: Number of articles per Domain and LT for Estonian.

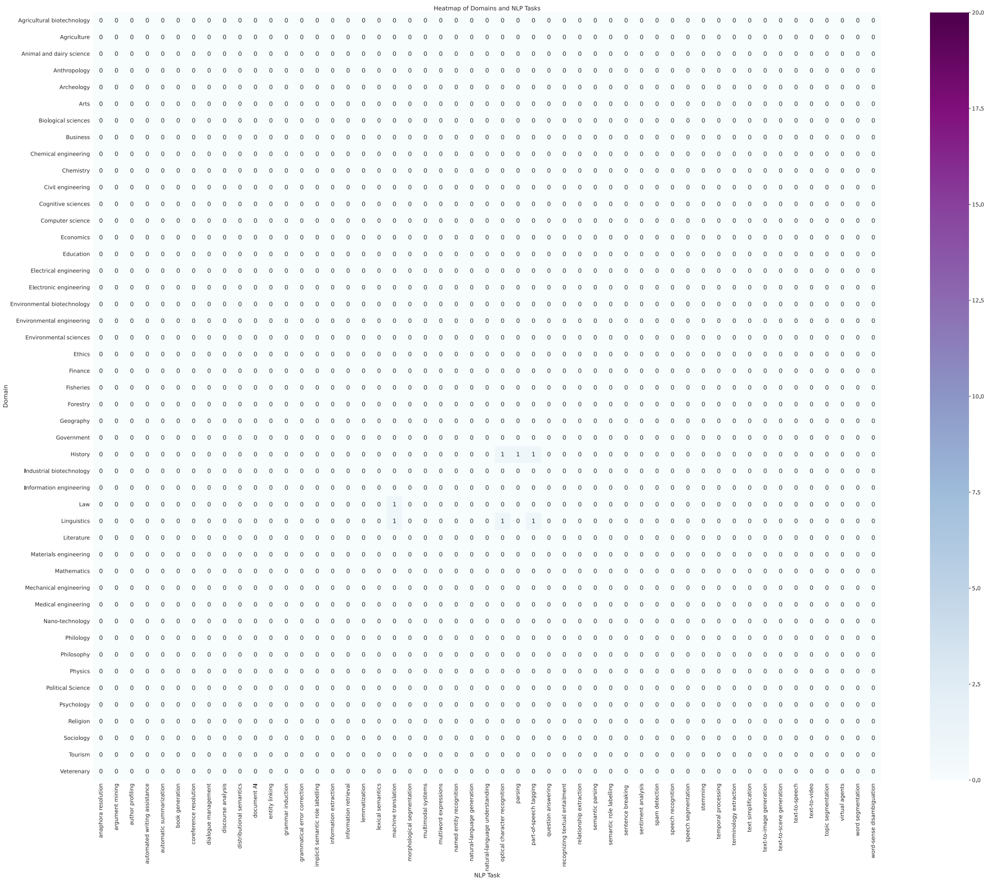

Figure 18: Number of articles per Domain and LT for Faroese.

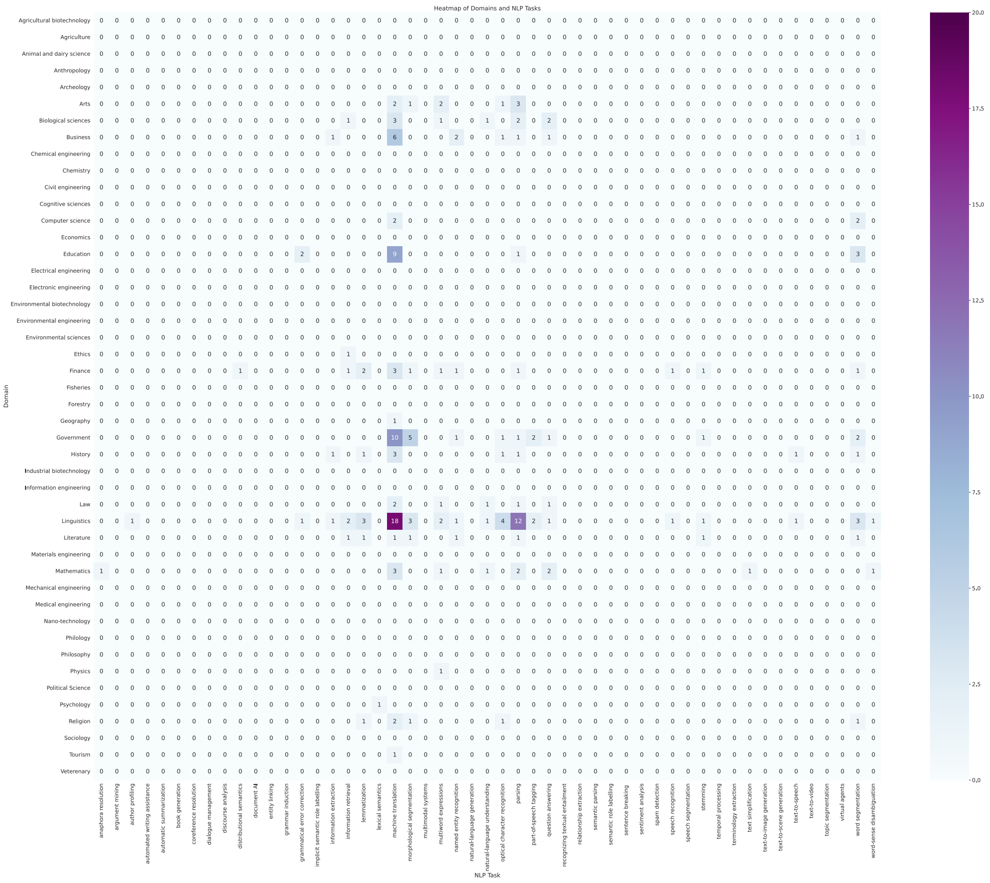

Figure 19: Number of articles per Domain and LT for Finnish.

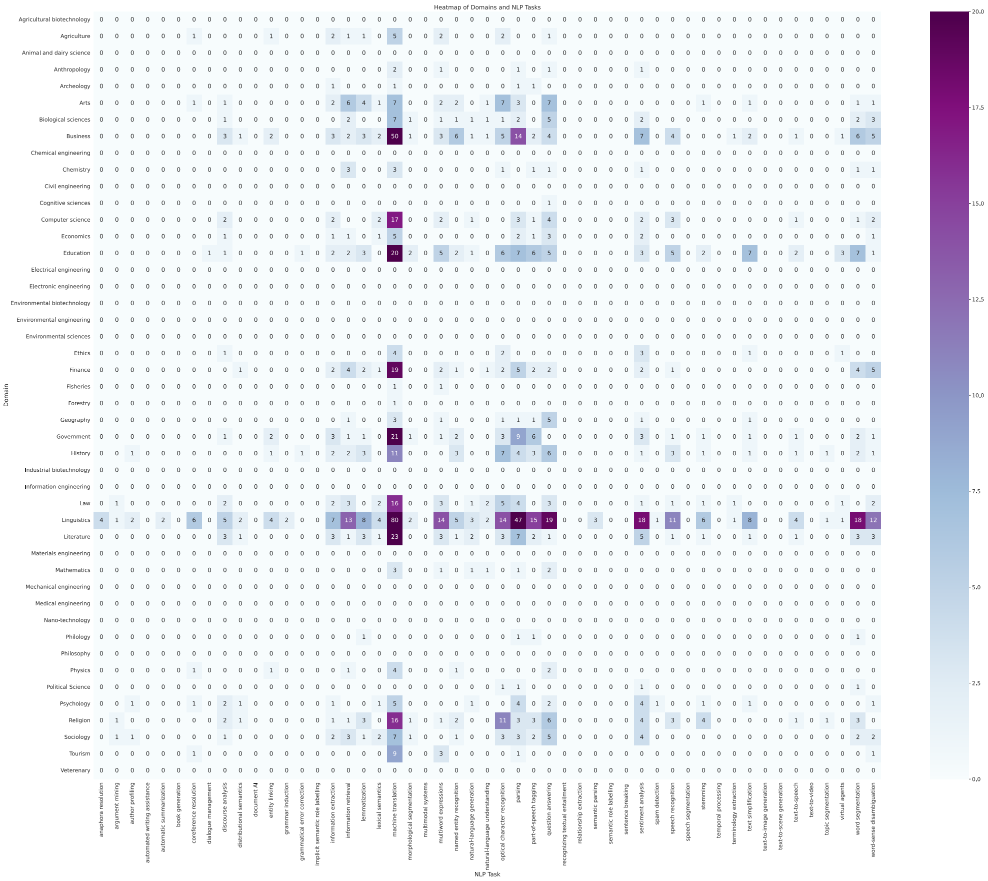

Figure 20: Number of articles per Domain and LT for French.

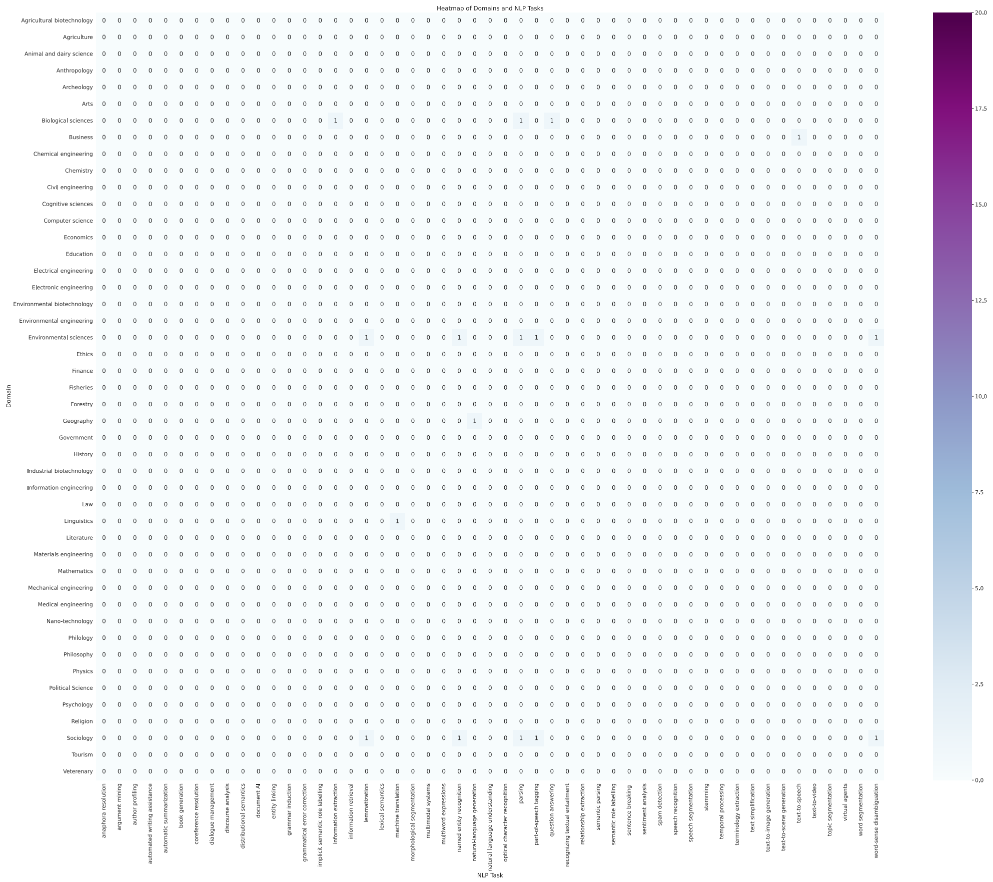

Figure 21: Number of articles per Domain and LT for Galician.

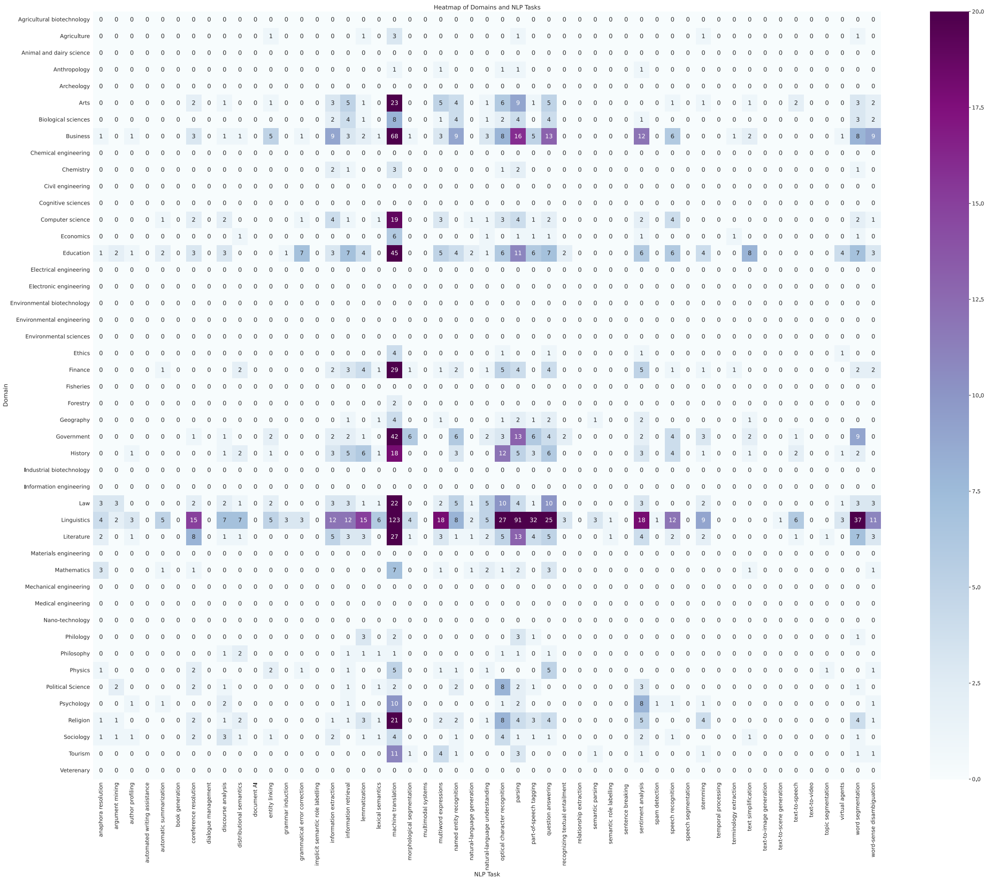

Figure 22: Number of articles per Domain and LT for German.

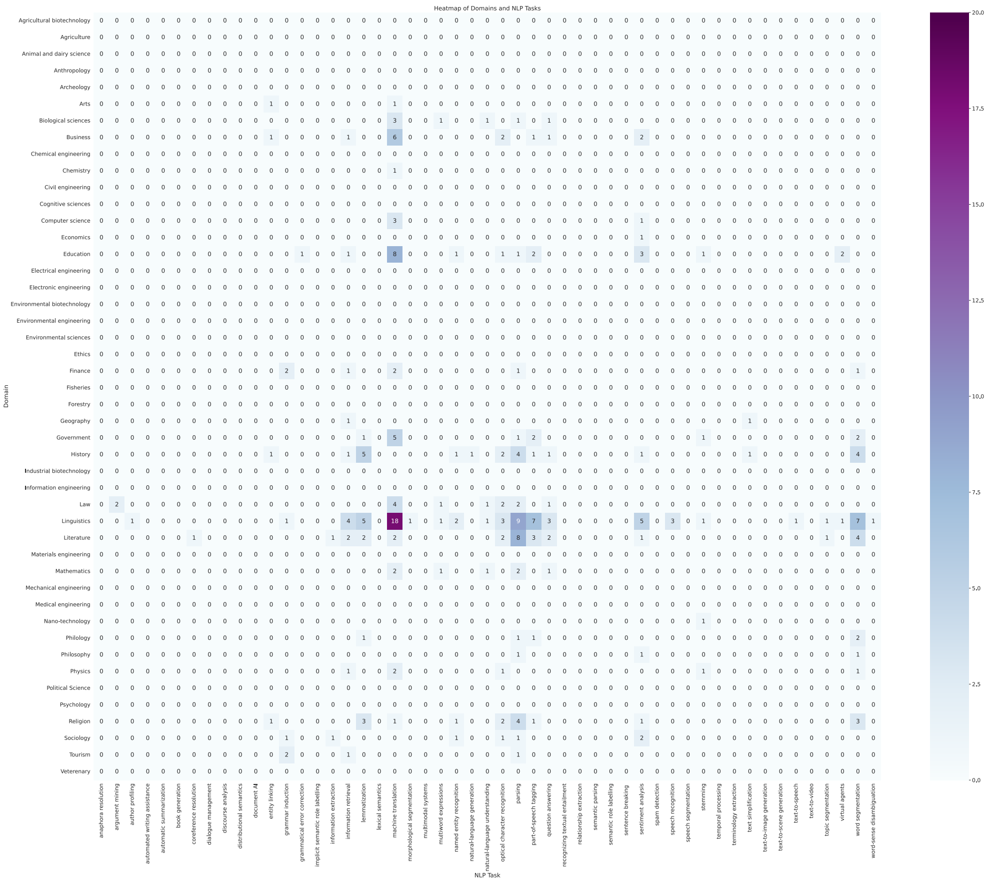

Figure 23: Number of articles per Domain and LT for Greek.

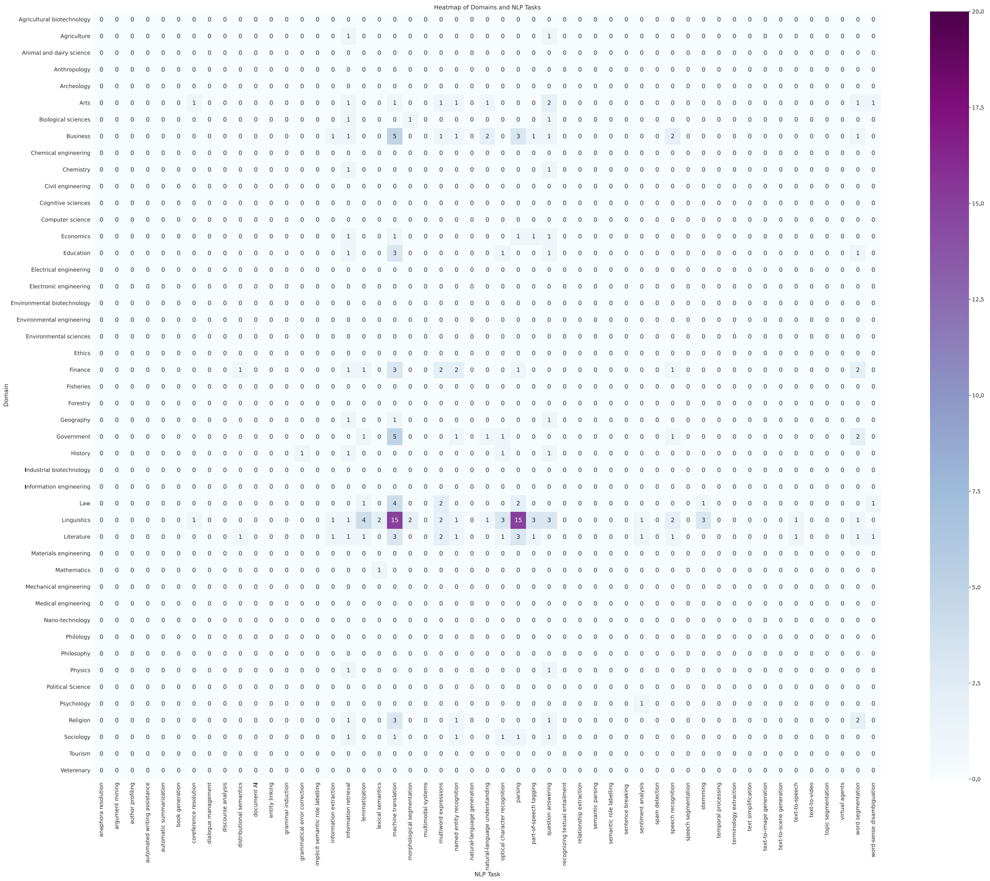

Figure 24: Number of articles per Domain and LT for Hungarian.

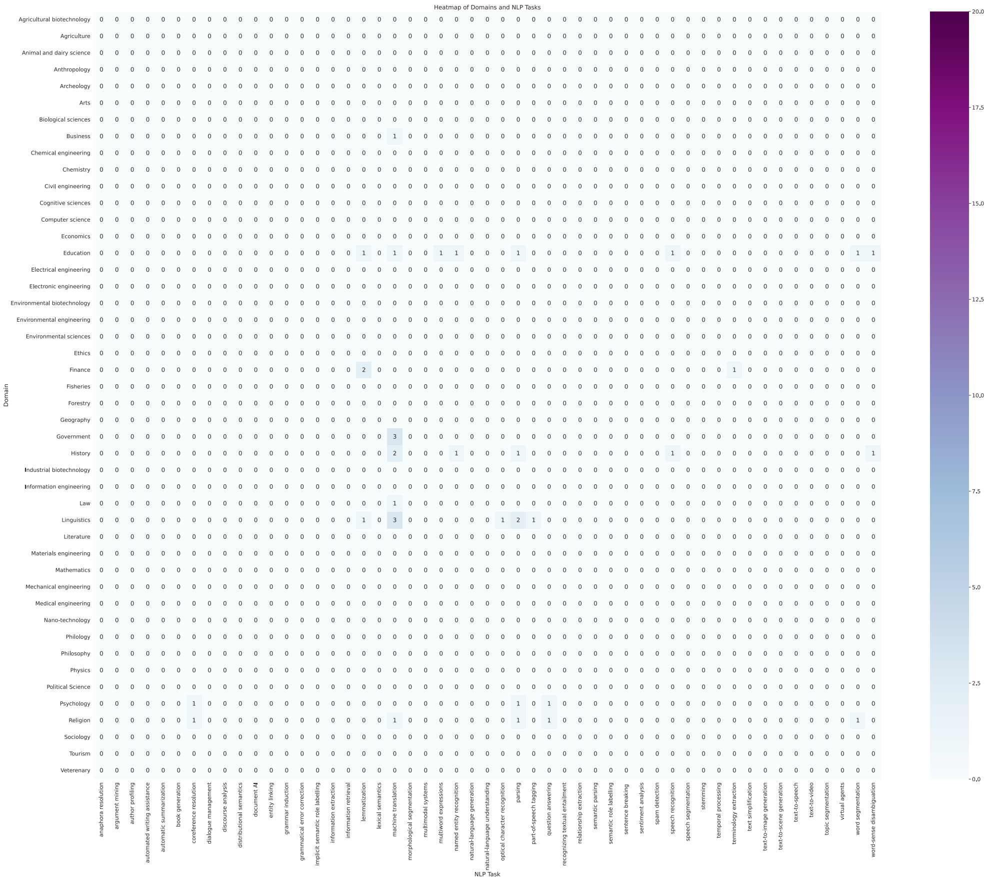

Figure 25: Number of articles per Domain and LT for Icelandic.

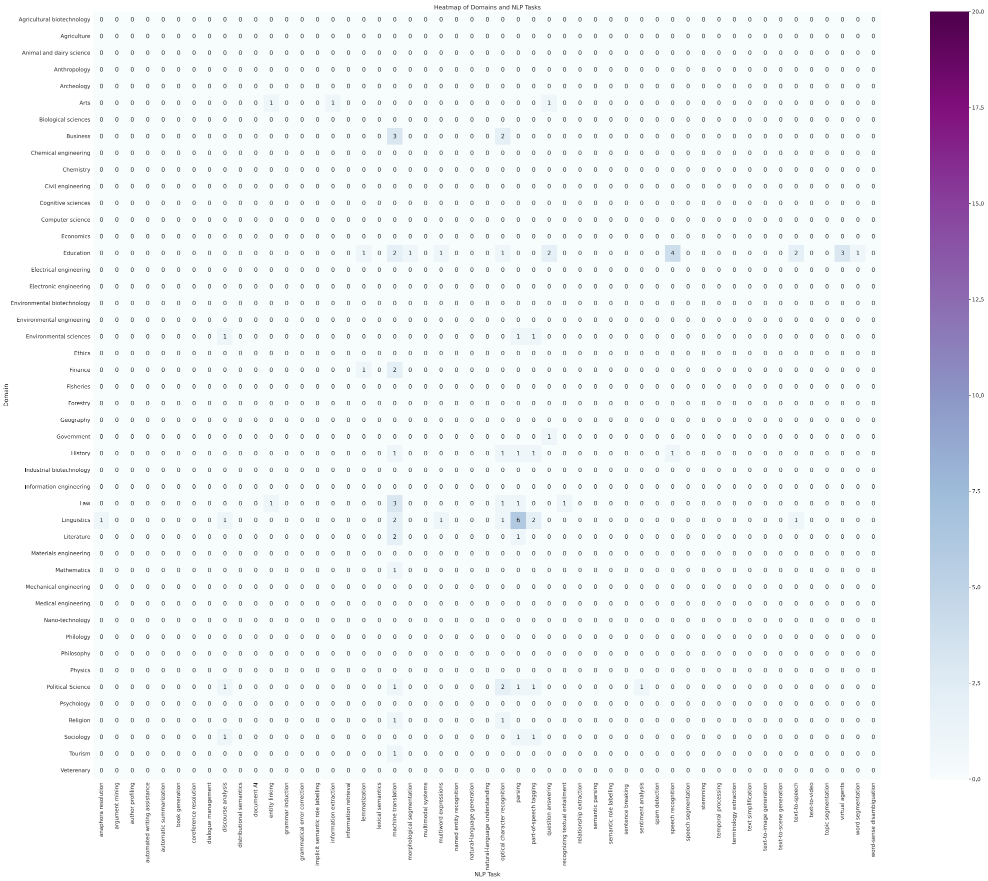

Figure 26: Number of articles per Domain and LT for Irish.

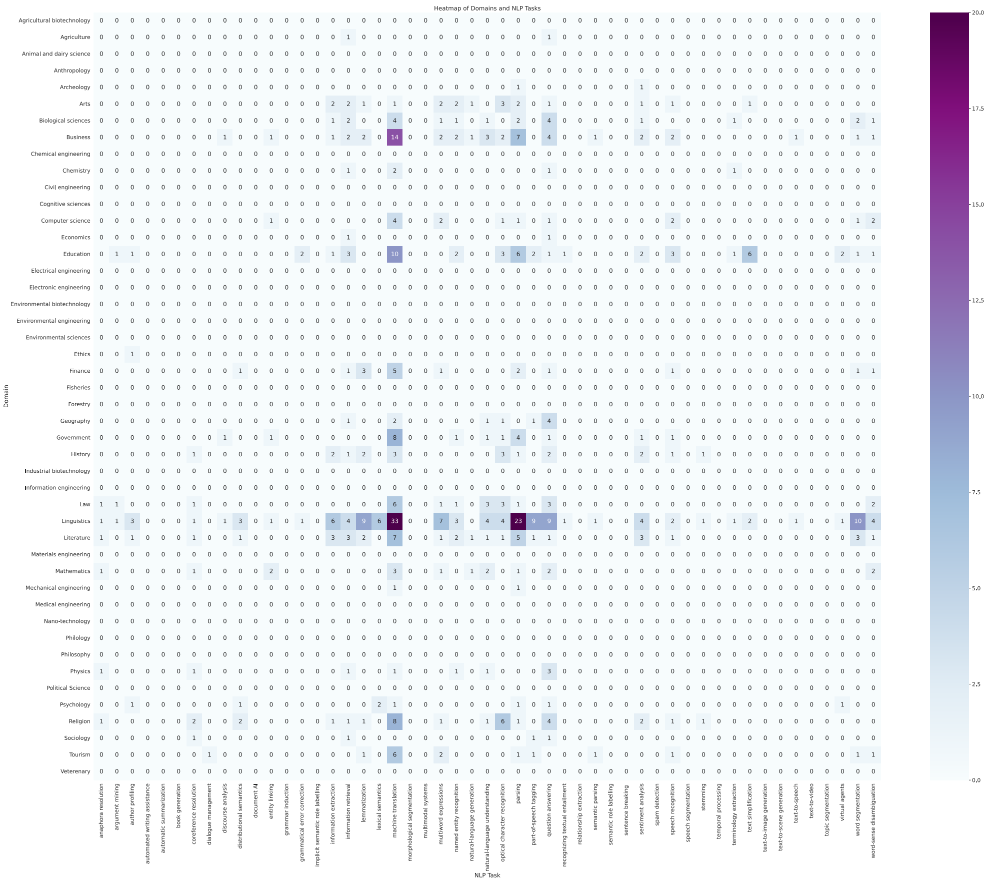

Figure 27: Number of articles per Domain and LT for Italian.

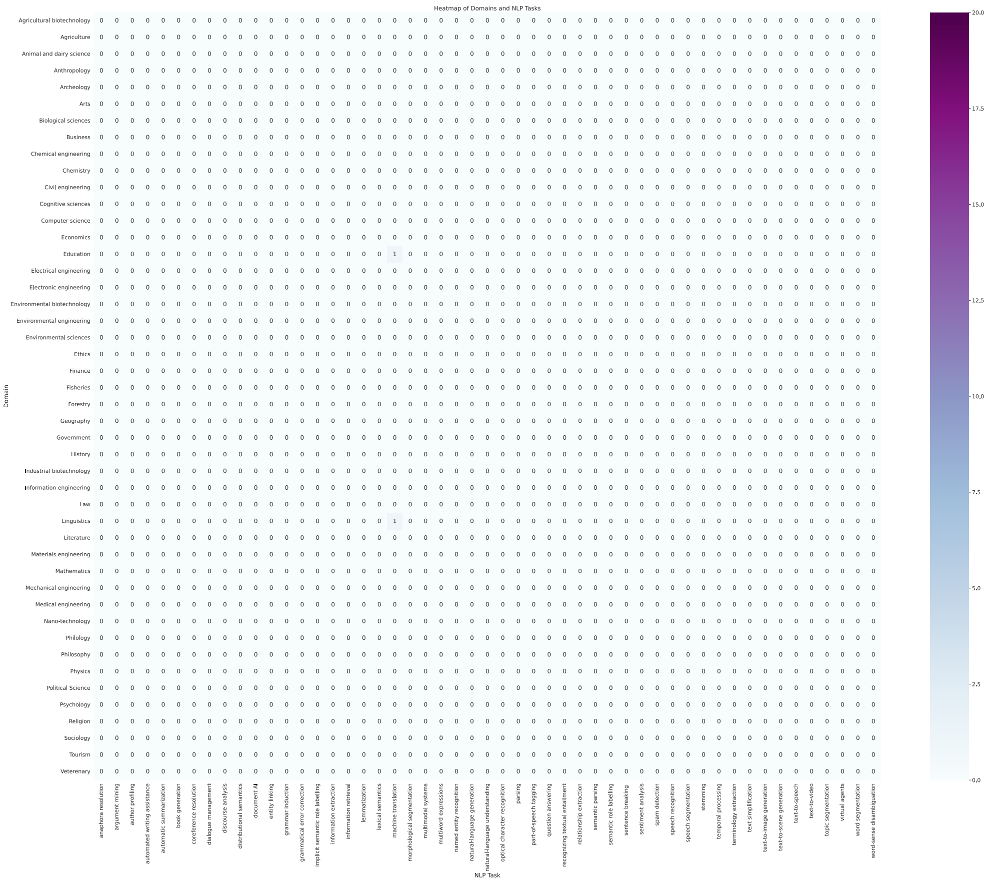

Figure 28: Number of articles per Domain and LT for Karelian.

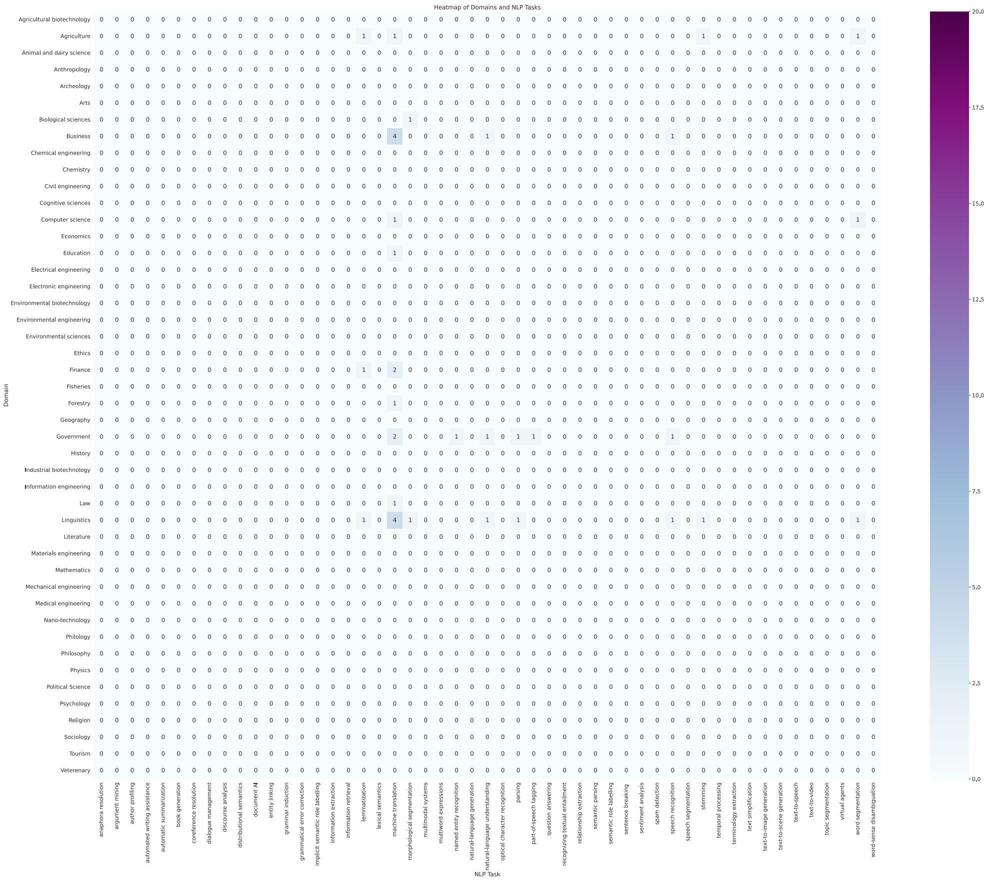

Figure 29: Number of articles per Domain and LT for Latvian.

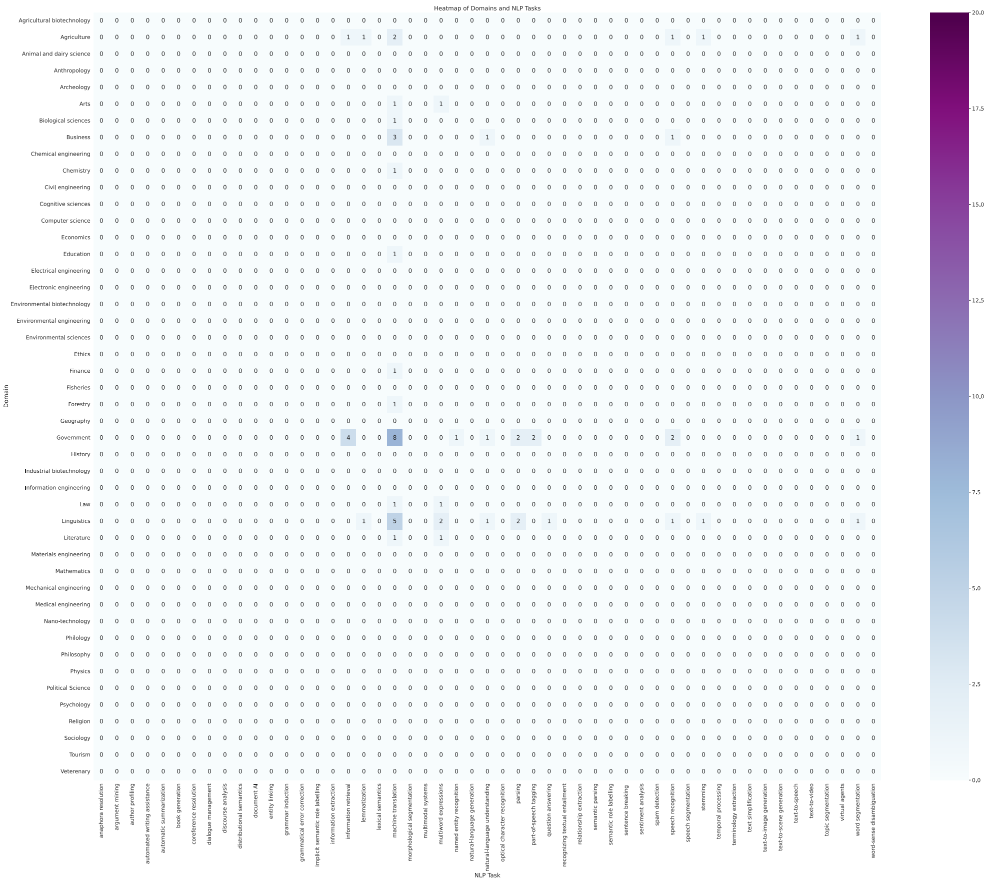

Figure 30: Number of articles per Domain and LT for Lithuanian.

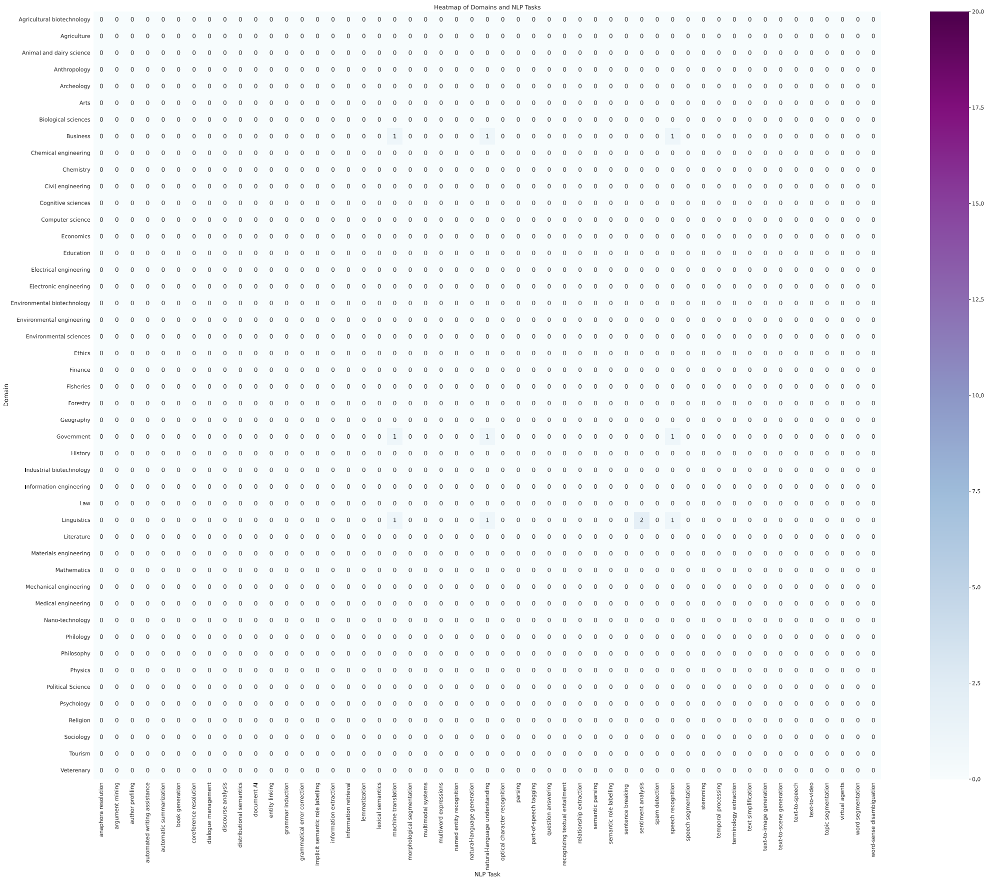

Figure 31: Number of articles per Domain and LT for Luxembourgish.

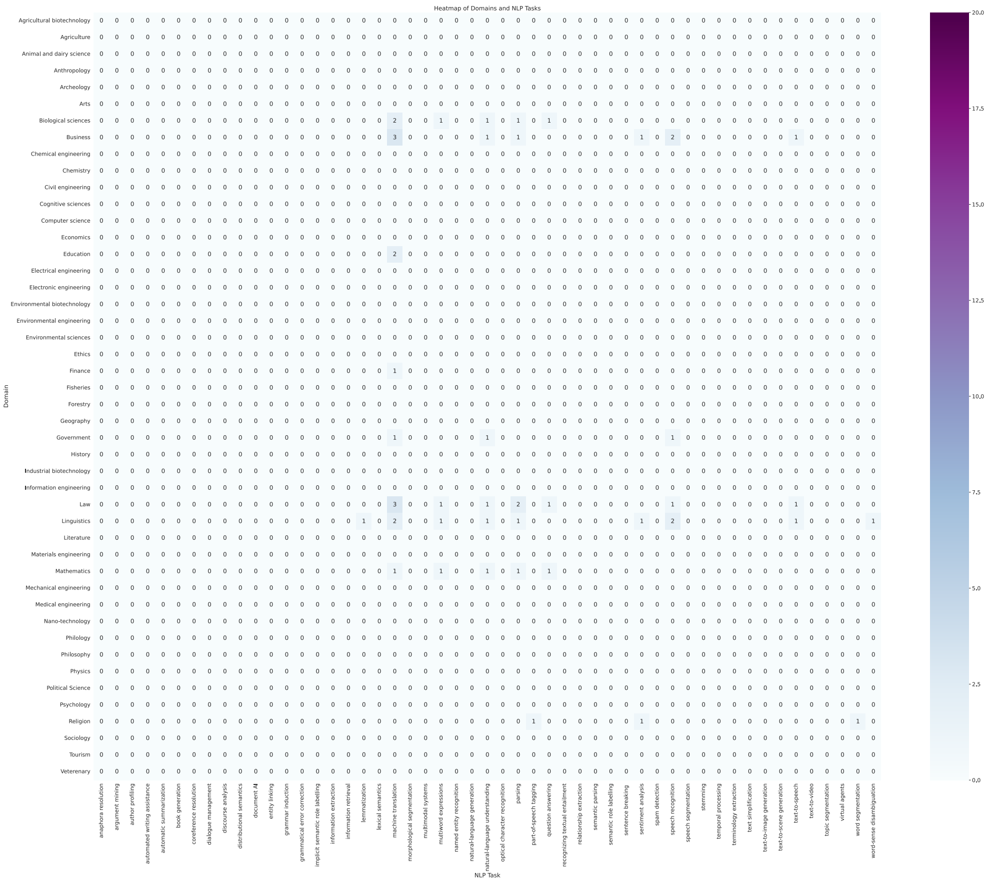

Figure 32: Number of articles per Domain and LT for Maltese.

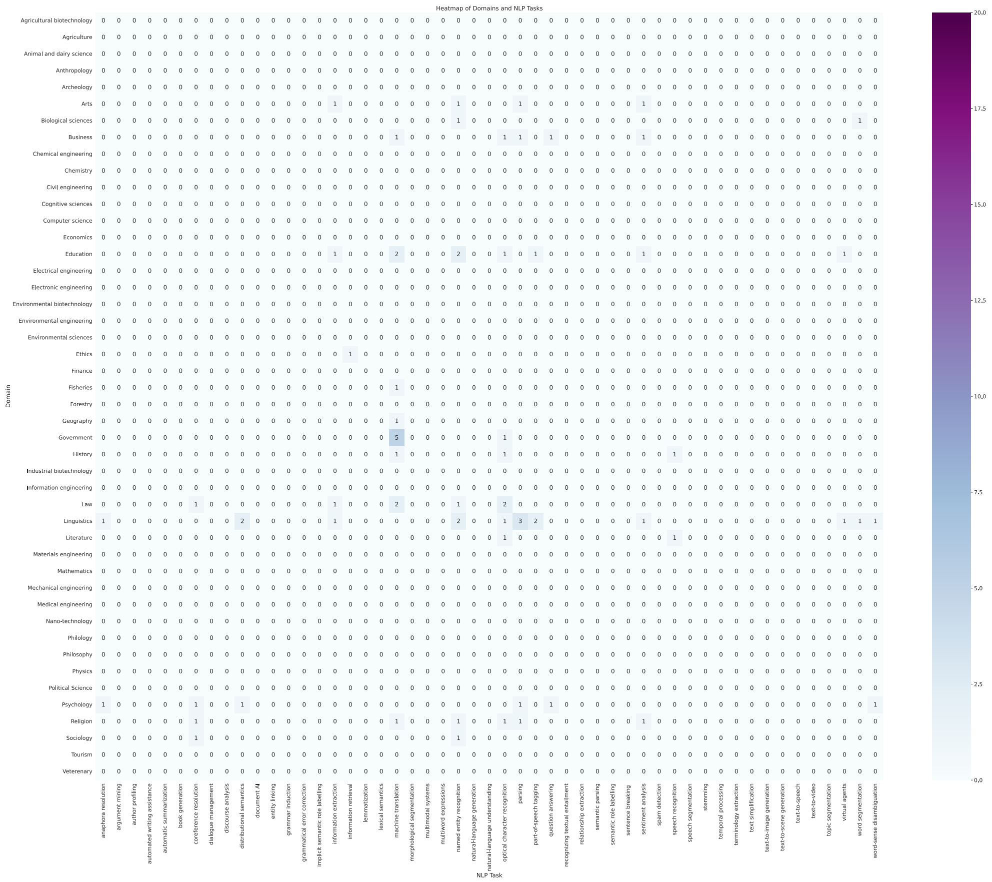

Figure 33: Number of articles per Domain and LT for Norwegian.

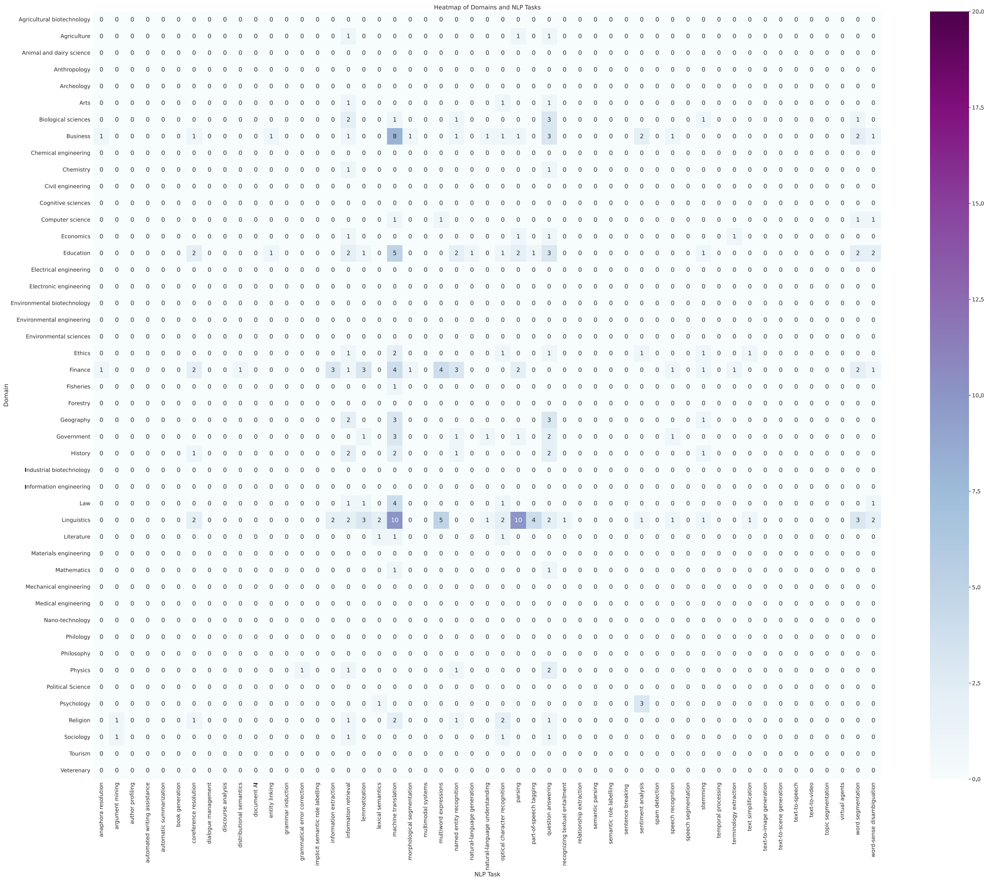

Figure 34: Number of articles per Domain and LT for Basque.

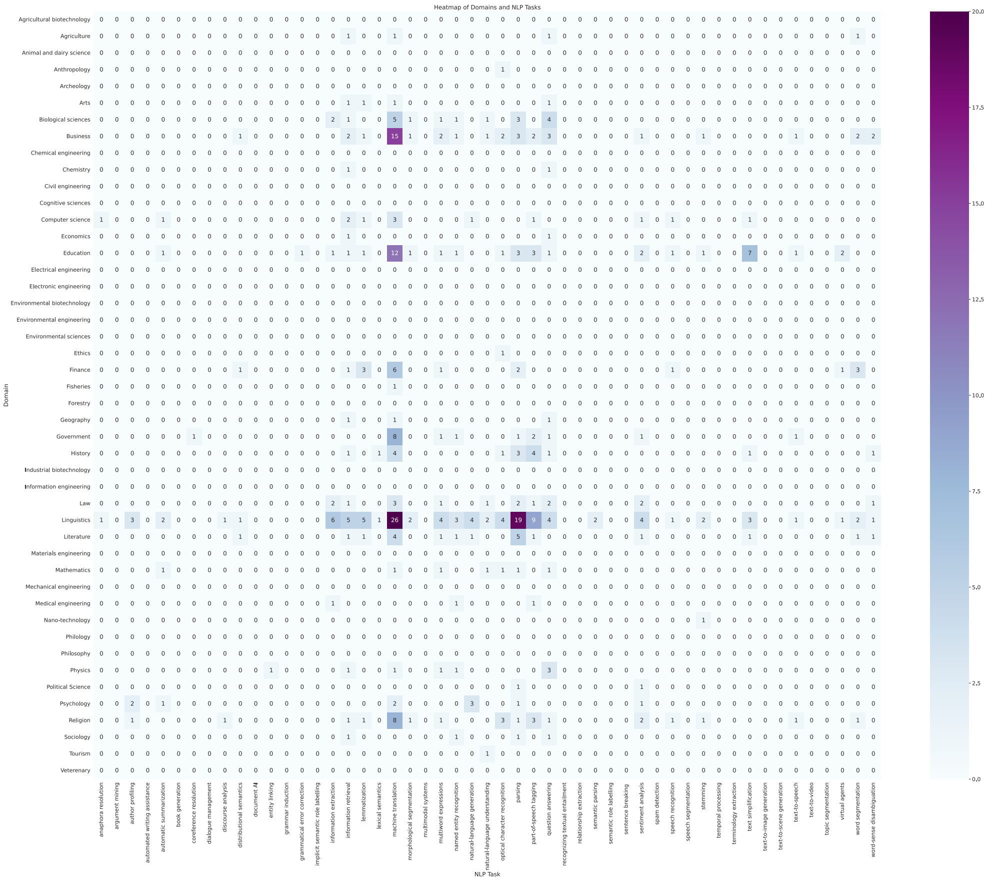

Figure 35: Number of articles per Domain and LT for Portuguese.

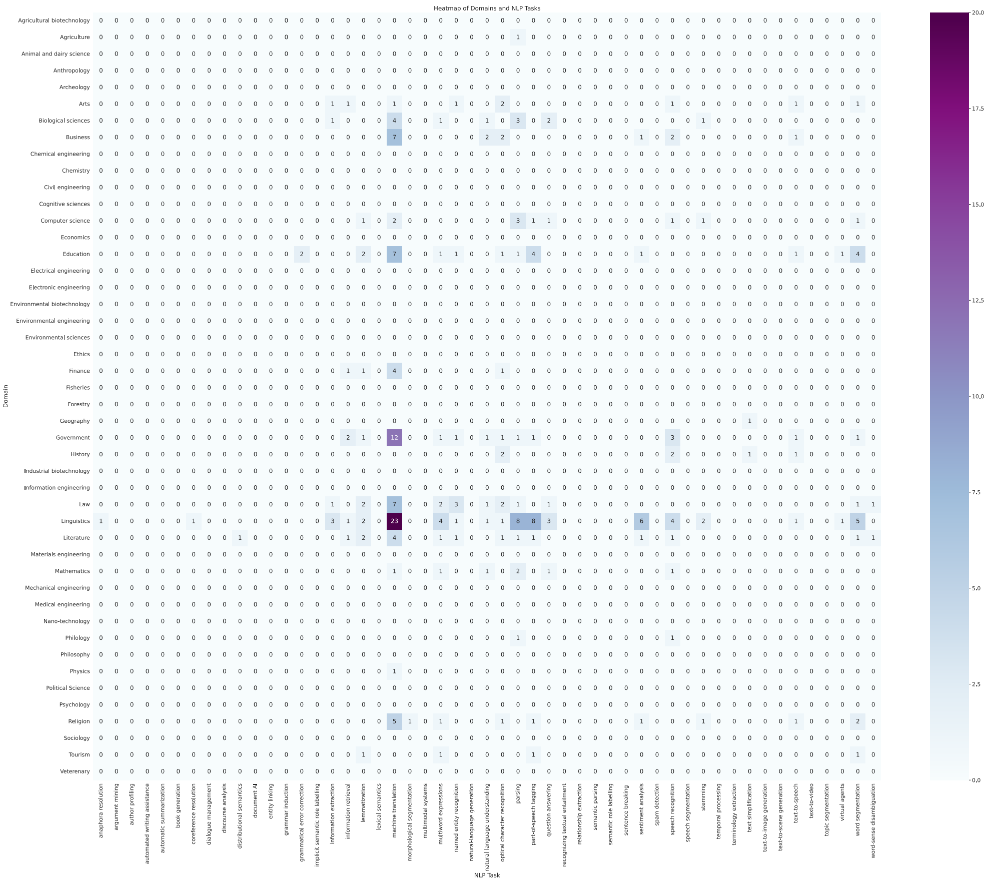

Figure 36: Number of articles per Domain and LT for Romani.

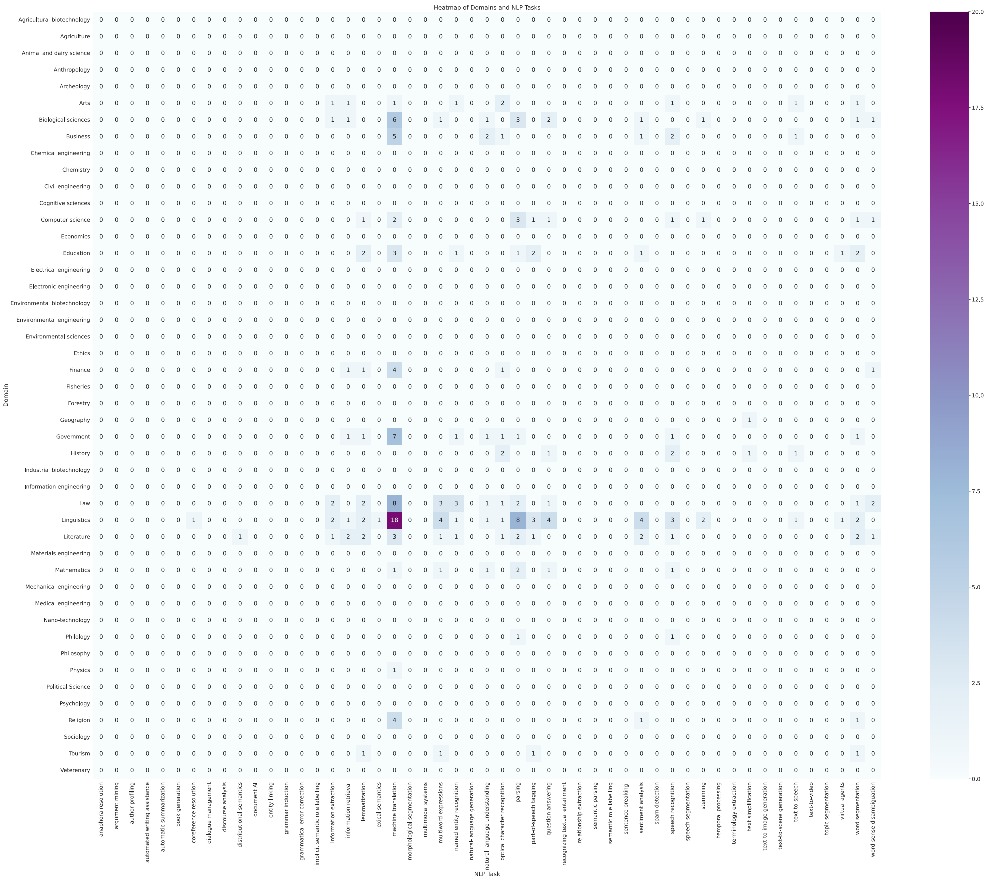

Figure 37: Number of articles per Domain and LT for Romanian.

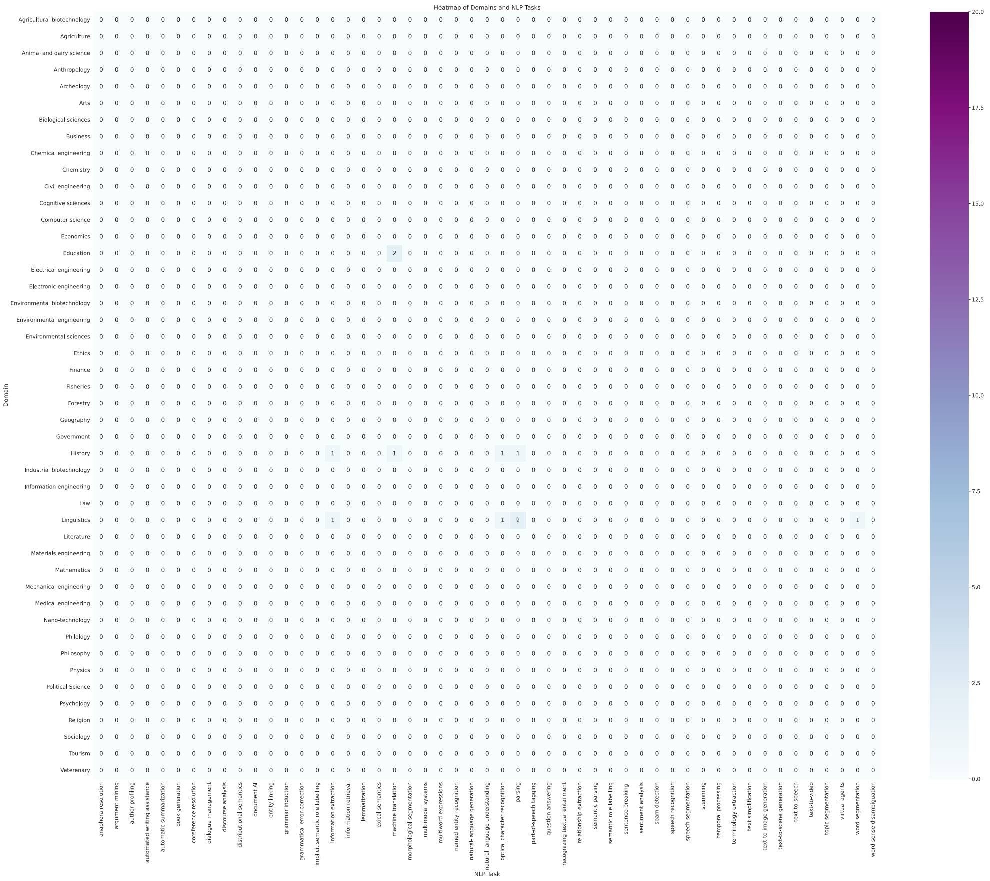

Figure 38: Number of articles per Domain and LT for Saami.

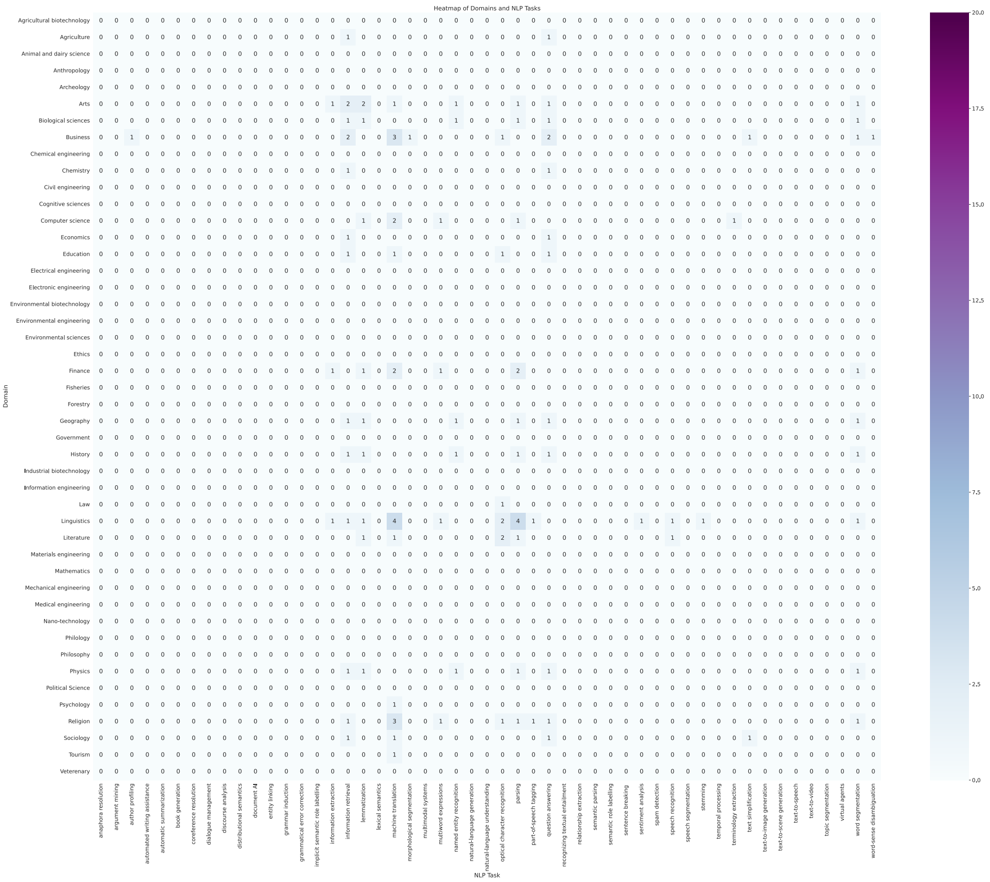

Figure 39: Number of articles per Domain and LT for Serbian.

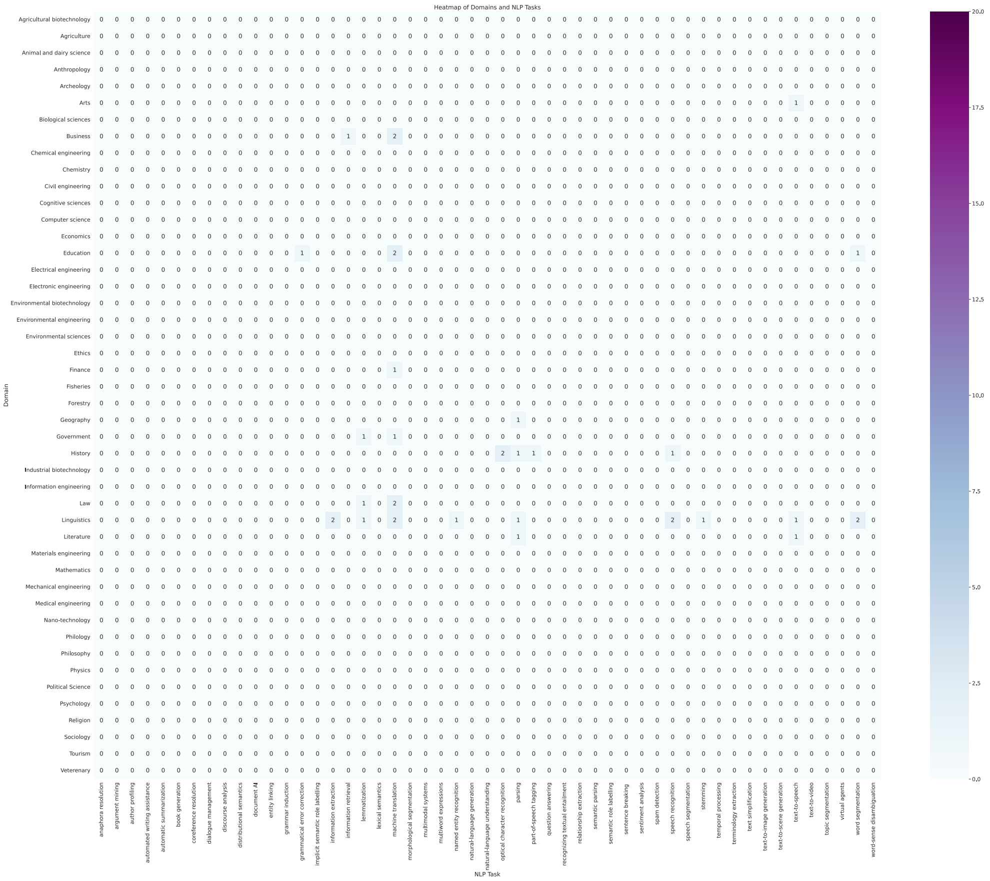

Figure 40: Number of articles per Domain and LT for Slovak.

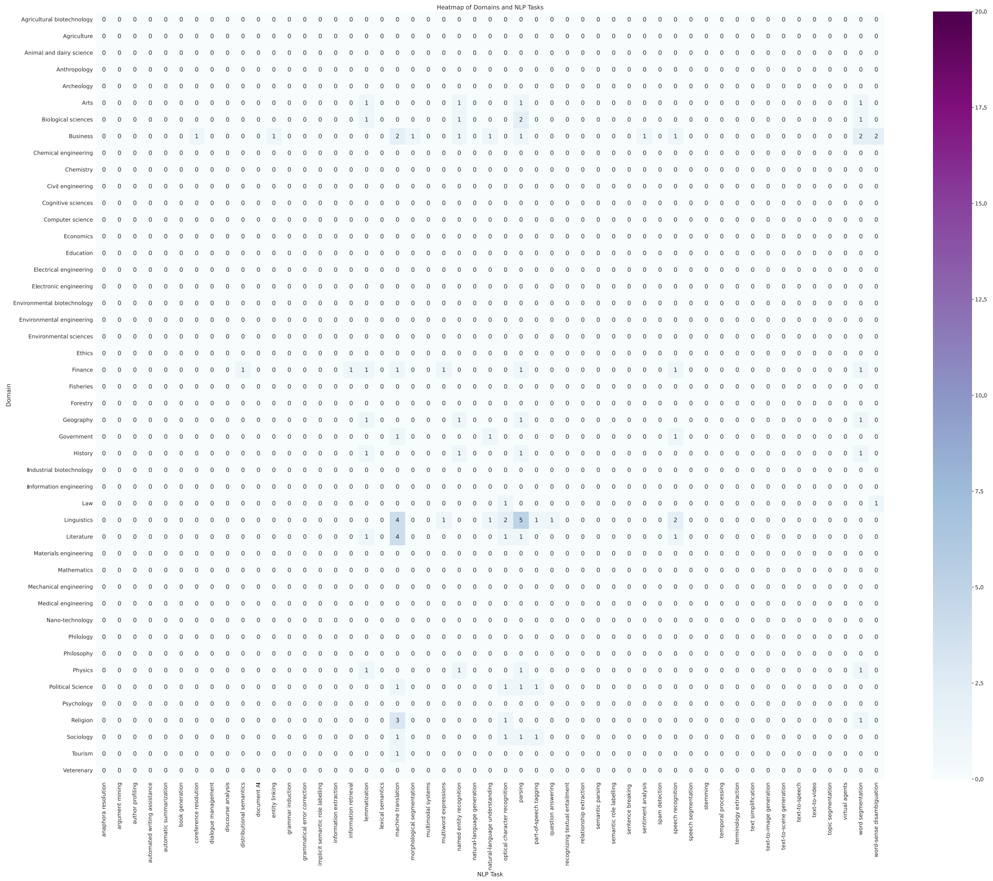

Figure 41: Number of articles per Domain and LT for Slovene.

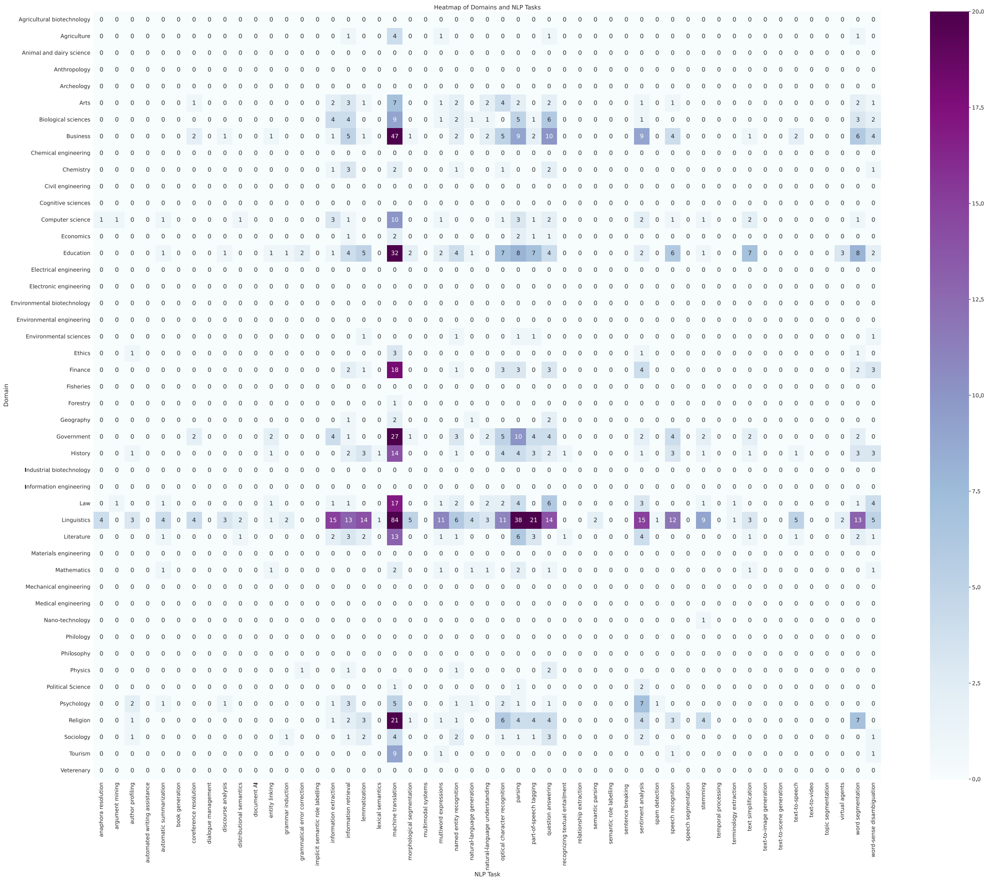

Figure 42: Number of articles per Domain and LT for Spanish.

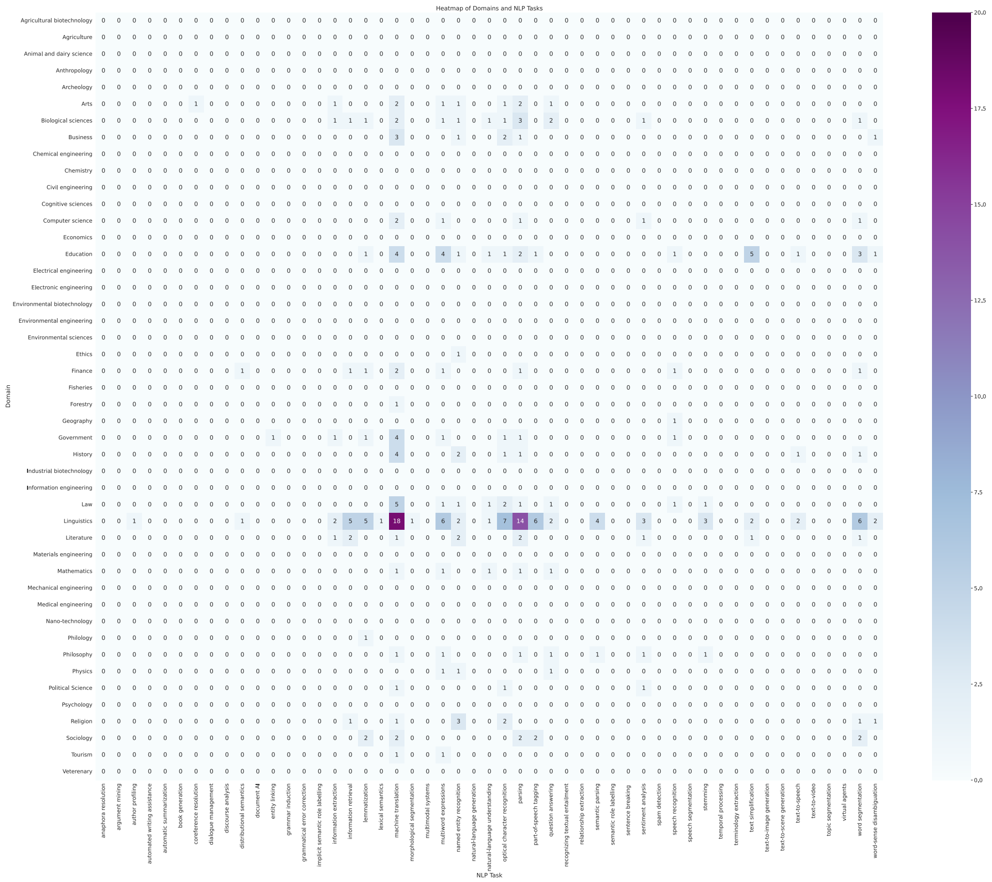

Figure 43: Number of articles per Domain and LT for Swedish.

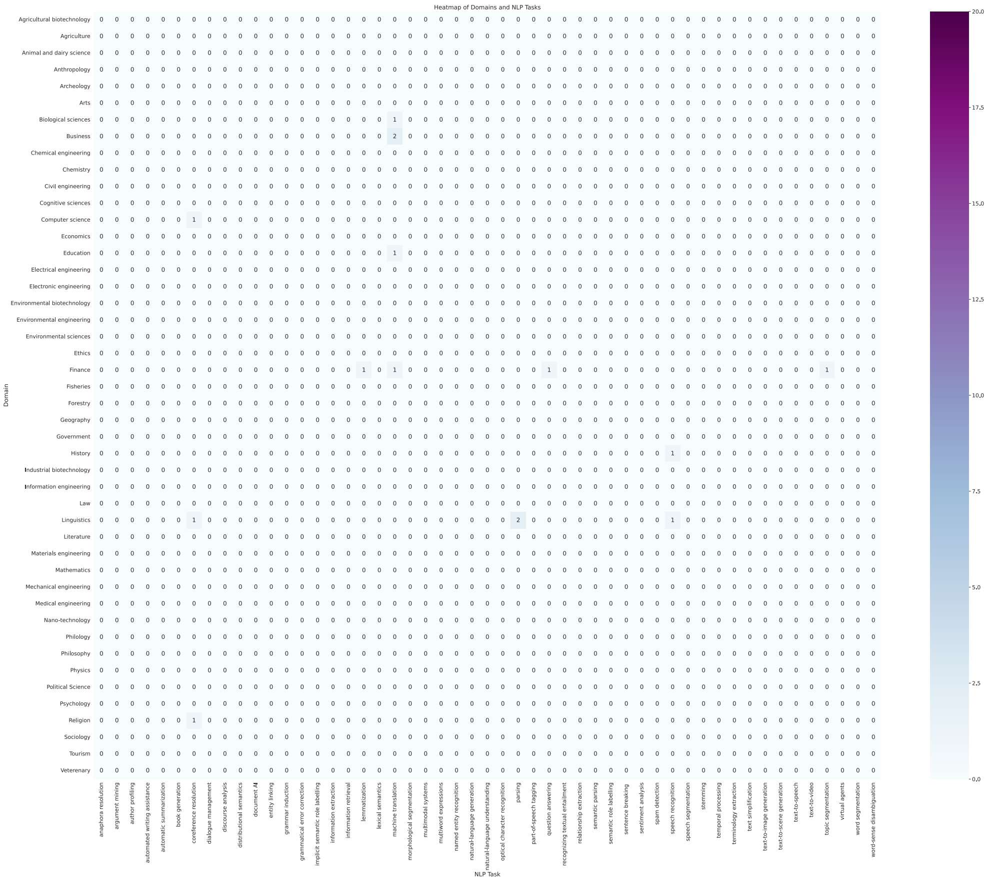

Figure 44: Number of articles per Domain and LT for Welsh.

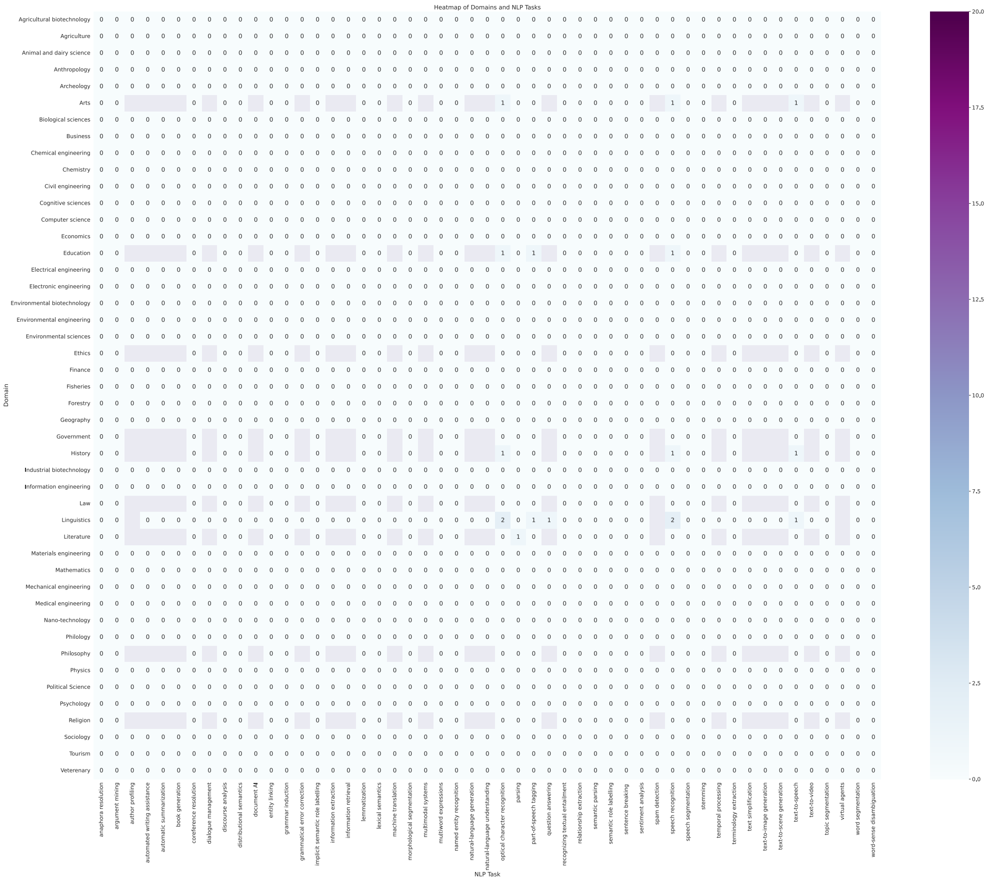

Figure 45: Number of articles per Domain and LT for Yiddish.

4.3 Discussion

In each part of the previous subsection, we presented the distribution regarding the usage of LT per domain for each language in our language set. As expected, the languages with more mentions in the ACL Anthology (Figure 3) have more complete heat-maps when compared to the others. However, even for them, it is possible to observe that not all domains and LT are completely developed in recent research papers.

By presenting the results for each language separately, our idea is to offer a clear view for the research community, enabling the identification of specific opportunities. It also allows the comparison between languages, thus, the more complete usage of LT by certain domains regarding some languages can inspire the scientific community to deploy them to the least developed ones.

It is also possible to notice that domains and NLP tasks that are more present in the selected papers (Figures 4 and 6) are the ones for which we can better extract information on how the domains are using different LT.

The main observed tendencies regarding the usage of NLP tasks by different domains are displayed in the heatmap presented in Figure 46.

As expected, Linguistics is the domain that has the highest number of associated NLP tasks as our data correspond to Computational Linguistics scientific work.

Besides that, it is possible to identify other domains with relatively high usage of different types of LT (i. e., at least one task with 20 articles or more): Arts, Biological sciences, Business, Computer science, Education, Ethics, Finance, Government, History, Law, Literature, Physics, Psychology, Religion, Sociology, and Tourism.

While some domains are covered by a large number of LT (e. g., Chemistry, Education, History, Religion, etc.), others rely on specific tasks. It is the case of the Ethics domain with a predominance of articles concerning sentiment analysis, machine translation, and question answering.

With Figure 46, it is possible to easily identify the domains and tasks with a lack of research that requires more attention from the NLP community.

5 Conclusions and Future Work

This report presented an overview of the usage of LT by different domains regarding a language set composed of 39 European languages as defined in the first ELE project.. The objective was to provide a snapshot of how domains and NLP tasks appear in research papers published by the LT scientific community from 2010 to 2022, allowing researchers to easily identify new opportunities for further development.

In the Methodology section, we presented the ACL Anthology Corpus collection, the chosen dataset for all the analysis. Then, we detailed how the lists concerning languages, domains, and NLP tasks were defined. Moreover, we described the dictionary-based approach which was used for the information extraction.

Regarding the results, first, we presented a separate overview regarding each dimension of this study. In the generated graphs, it is possible to identify the languages, domains, and LT that are most cited in the data-set and the ones which are underrepresented.

Then, we generated, for each language, heat-maps that detail the usage of the selected NLP tasks by the different domains. We completed this language-specific examination with a general overview of the distribution of LT by domain. The domains which clearly present a broader usage of LT are: Arts, Biological sciences, Business, Computer science, Education, Ethics, Finance, Government, History, Law, Literature, Physics, Psychology, Religion, Sociology, and Tourism. However, while a few domains use several NLP tasks, others focus on a small number of them.

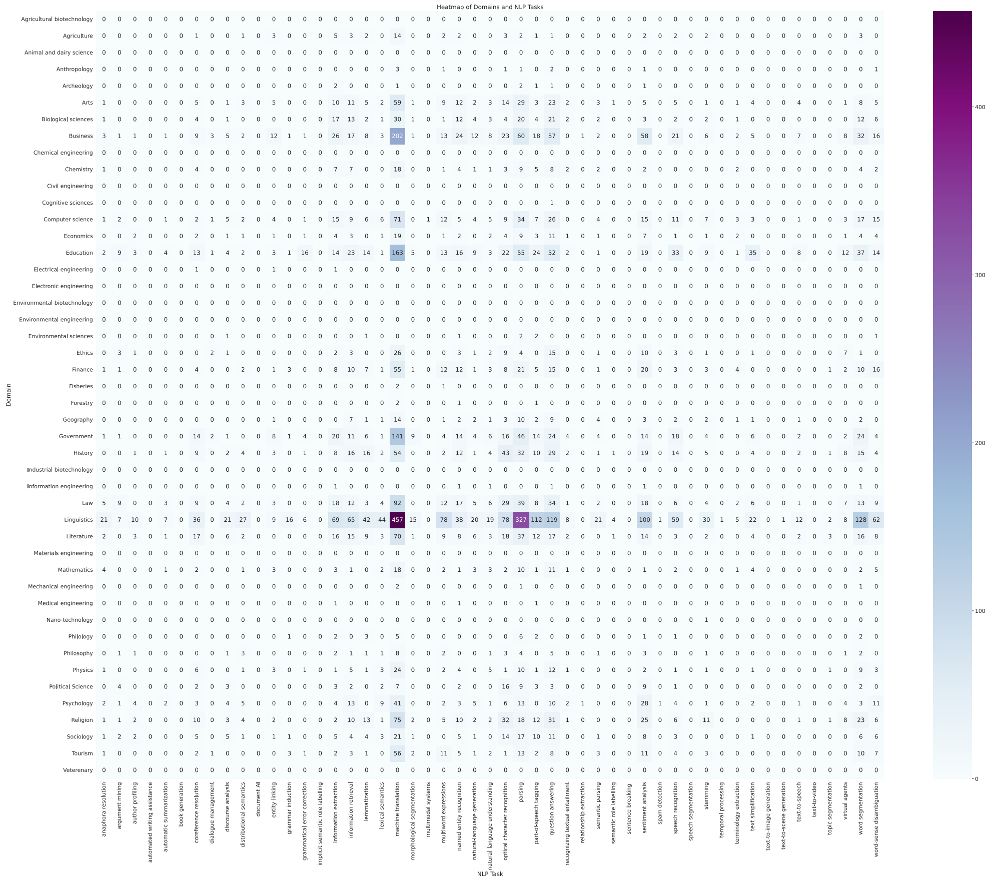

Figure 46: Number of articles per Domain and LT for all languages considered in this report

On the other hand, the domains which are not mentioned in the texts are: Agricultural biotechnology, Animal and dairy science, Chemical engineering, Civil engineering, Electronic engineering, Environmental biotechnology, Environmental engineering, Industrial biotechnology, Materials engineering, Veterinary.

For better language equality, ideally, the scientific community should focus on the under-represented languages and on the specific NLP tasks and domains in the heat-maps which present 0 articles.

It is important to mention that, although presenting a clear overview of the situation regarding domains and tasks, this study has some biases that need to be taken into account. The dictionary-based approach was chosen as it was the most adapted one for this analysis regarding time and resources, however, as it considers only the exact terms on the defined lists, it may miss and omit some information.

In this report, we analysed only LT papers written in English, while the overall picture would be more complete if national LT conferences and their production were considered. This could be a useful topic for a future iteration of such survey.

Furthermore, this study considers only the research published in the ACL Anthology. Thus, it focuses on how the NLP community deal with the languages, domains, and LT. As a perspective for a future ELE project, a complementary study could be conducted considering other research databases such as Web of Science[10] and Scopus[11]. In these resources, each article is manually classified in terms of domains, thus, it could provide valuable information on how different LT are being used in scientific research being done by professionals from other domains and other scientific fields.

References

ACL - Association for Computational Linguistics. https://www.aclweb.org/. Accessed: 2023-03-20.

GROBID. https://github.com/kermitt2/grobid, 2008–2023. Accessed: 2023-01-30.

Rodrigo Agerri, Eneko Agirre, Itziar Aldabe, Nora Aranberri, Jose Maria Arriola, Aitziber Atutxa, Gorka Azkune, Arantza Casillas, Ainara Estarrona, Aritz Farwell, et al. European language equality d1. 2: Report on the state of the art in language technology and language-centric ai, september 2021. 2021.

Daniel Gildea, Min-Yen Kan, Nitin Madnani, Christoph Teichmann, and Martín Villalba. The ACL Anthology: Current state and future directions. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 23–28, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2504. URL https://aclanthology.org/W18-2504.

IBM. What is natural language processing (nlp)? https://www.ibm.com/topics/natural-language-processing. Accessed: 2023-01-20.

Nitin Indurkhya and Fred J Damerau. *Handbook of natural language processing*. Chapman and Hall/CRC, 2010.

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

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

Frascati Manual et al. Guidelines for collecting and reporting data on research and experimental development. 2015. URL https://www.oecd.org/innovation/frascati-manual-2015-9789264239012-en.htm.

[10] https://clarivate.com/webofsciencegroup/solutions/web-of-science/

[11] https://www.scopus.com/

Maite Melero, Pablo Peñarrubia, David Cabestany, Blanca C. Figueras, Mar Rodríguez, and Marta Villegas. European language equality - d1.32 - report on the spanish language, 2022.

Ruslan Mitkov. *The Oxford handbook of computational linguistics*. Oxford University Press, 2022.

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

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

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

Georg Rehm and Stefanie Hegele. Language technology for multilingual europe: An analysis of a large-scale survey regarding challenges, demands, gaps and needs. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Shaurya Rohatgi. Acl anthology corpus with full text. Github, 2022. URL https://github.com/shauryr/ACL-anthology-corpus. Accessed: 2023-01-10.

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

Wikipedia. Natural language processing. https://en.wikipedia.org/wiki/Natural_language_processing. Accessed: 2023-01-20.