Visualizing Language Diversity in the ACL Anthology Corpus

Annebeth Buis*
University of Colorado Boulder

Abstract

This project describes two visualizations that were created to visualize language diversity in the ACL Anthology Reference Corpus. In the first visualization - a stacked area chart - the number of paper over time is shown for the 10 most frequent languages. The second visualization - a categorical bubble chart - encodes relationships between papers and research topics, aggregated over time.

1 Introduction

Natural Language Processing (NLP) is an interdisciplinary field that concerns natural human languages. Approximately 7000 languages exist in the world right now, but for almost all of those languages little to no data is available. Because of that, NLP has historically focused on big, well-known and mostly Indo-European languages. Especially English has had a privileged role in NLP: most work is done on English data and it is also usually considered the *default* language.

In this project, I am looking at languages mentioned in proceedings papers of the Association for Computational Linguistics (ACL). My goal is to visualize language diversity in ACL papers and to give insight to how this has changed over time. I will also look at how language relates to research topics. For example: does the Semantic Role Labeling community more often work on Arabic? Is dependency parsing more popular for Czech?

For this purpose, I created two visualizations. The first shows language diversity over time through a stacked area chart. The second captures the relationship between languages and research topics in a categorical bubble chart. This projects hopes to stress the current under-representation of languages besides English. It also gives an insight into a field of research and adds to existing work on the ACL Anthology.

2 Related Work

The motivation for this project comes from recent developments in the computational linguistics community. Researchers noticed a pattern of more papers being submitted with no mention of which language the paper is solving a problem for. Historically, NLP research has focused mostly on English and often “No language”-papers simply assume that English is the “default” language. Recently, Emily Bender coined the “Bender rule”: you have to name the language you are working on [2]. The ACL Anthology Corpus only contains papers up to 2010, which makes it interesting to see how this habit has developed over time.

In related work on the ACL corpus, [1] aim to create an author-centered computational history of the field. Based on the corpus data, they identify 4 distinct periods in NLP research. They also propose a method to obtain research topics or subfields directly from the corpus data. By running LDA topic models over the entire corpus, they produce 100 generative topics. These topics were then hand-labeled by experts in the field. In this project, this list of topics will be used as the basis for labeling the papers with research topics.

To my best knowledge, there have been no attempts to visualize the ACL data, except for network analysis of citations (e.g., [7]). Previous work on visualizing relationships between journal papers has shown the potential of cross-maps (or correlograms/bubble charts). This has inspired me for the choice of my second visualization for this project.

3 Design of the Visualizations

3.1 ACL Anthology

The ACL Anthology Reference Corpus [3] is available for download online.[1]. For this project, both the ACL XML Metadata and the OmniPage OCR XML were used. The metadata contains all information about title, year and authors and the OCR’ed XML files contain all the text from the original proceedings PDF.

The corpus contains papers from many ACL conferences, including EMNLP, COLING, etc. However, to limit the dataset, in this paper I only worked on papers published at the main ACL event. The corpus includes all papers from 1979 to 2010.

3.2 Preprocessing

The preprocessing for this experiment was substantial. As a first step, the metadata needed to be combined with the paper text, to obtain the complete information on each paper. I wrote a script that extracted each word from the OCR’ed XML files and saved it with the correct paper information.

In the next step, I used a simple heuristic to determine whether a paper was *about* or *working on* a certain language. I combined the current language inventory from the UniMorph project [5] and the Universal Dependencies project [6], resulting in a list of 146 unique languages. These projects are both known for their wide language coverage, so it is very unlikely that a paper discusses a paper that is not covered in either.

For each paper, I checked whether a language from the list occurs in the text. Naturally, this is a heuristic and not perfectly accurate. However, it is unlikely for researchers to mention languages that are not further discussed in the paper. Also, some work in NLP is multi-lingual, so papers can be associated with multiple languages.

To determine a research topic for each paper, I used the list that was created by applying LDA topic modeling to the corpus (discussed in the previous section). Since the main goal of this project was to create visualizations, I went with a simple heuristic approach and looked for the exact research topic “name” in the text. For example, a research topic from the list is *Machine Translation*, so I looked for instances of that exact term in the text. Each paper should fall under one main research topic, so I picked the topic with most occurrences for each paper. Note that this is a very rudimentary approach and that it has likely affected the quality of the dataset. In the future, it would be beneficial to take a Machine Learning approach to classifying the papers.

Besides the preprocessing of the corpus files, I also created several scripts to count and create the correct input datasets for my visualizations.

3.3 Stacked area chart for language diversity

A stacked area chart consists of several area charts for different categories, stacked on top of each other. By stacking, they present an

*e-mail: anne.buis@colorado.edu

[1] `https://acl-arc.comp.nus.edu.sg/`

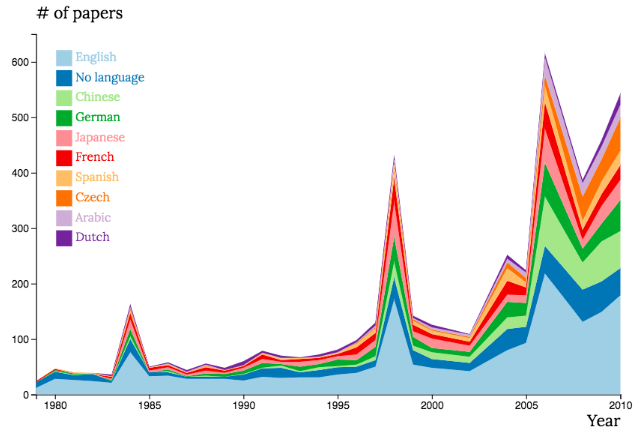

Figure 1: Stacked area chart showing language diversity in the corpus from 1979 to 2010.

aggregate of the information, which is intuitive when all categories together represent a whole. Traditionally, stacked graphs have been known to be hard to read, because it can be difficult to interpret trends that have been stacked on top of each other [4].

In my particular use-case, however, the user is probably not interested in the exact number of papers that was published in a certain year or the exact difference between two languages. The visualization allows for encoding multiple layers of information at once: the overall increase in publications over the years, the big visual difference between English and the other languages in the top 10 and also the increasing popularity of some languages. That same information would not have been clear from multiple line graphs or a similar representation.

Visualization 1 shows that even though English still is the main language researchers in the NLP community work on, the increase in productivity has also lead to an increase of language diversity. More disturbing, from a linguist's perspective, is that there are also more papers in which no language is mentioned at all. This points again to the Bender rule [2] (discussed in related works): researchers should name the language that they are working on.

The visualization is implemented with interaction: on hovering over the legend, the associated area is highlighted. This is helpful for tracking the width for one individual language over time, allowing the user to focus on another task.

3.4 Categorical bubble chart for research topics

Visualization 2 is a correlogram-bubble chart hybrid. In a correlogram, traditionally, the represented circles are correlations. Bubble charts often contain at least one quantitative axis. In this visualization, the bubbles represent the number of papers that exists for a given language-research topic combination.

During the prototyping phase of this project, I considered several other visualizations for this dataset, including a sankey and a bipartite graph. However, since almost all languages are connected to almost all research topics, those result in a tangle of lines from which it is very hard to get a visual gist. This visualization is extremely clean and bigger bubbles pop-out and can visually cluster.

Clustering patterns would be more obvious had their been more variation in the data. However, in practice we see that for most research topics, bubble sizes are consistently small across languages. Only a few cases stand out, such as Dependency Parsing for Czech - which is a well known language-research topic combination.

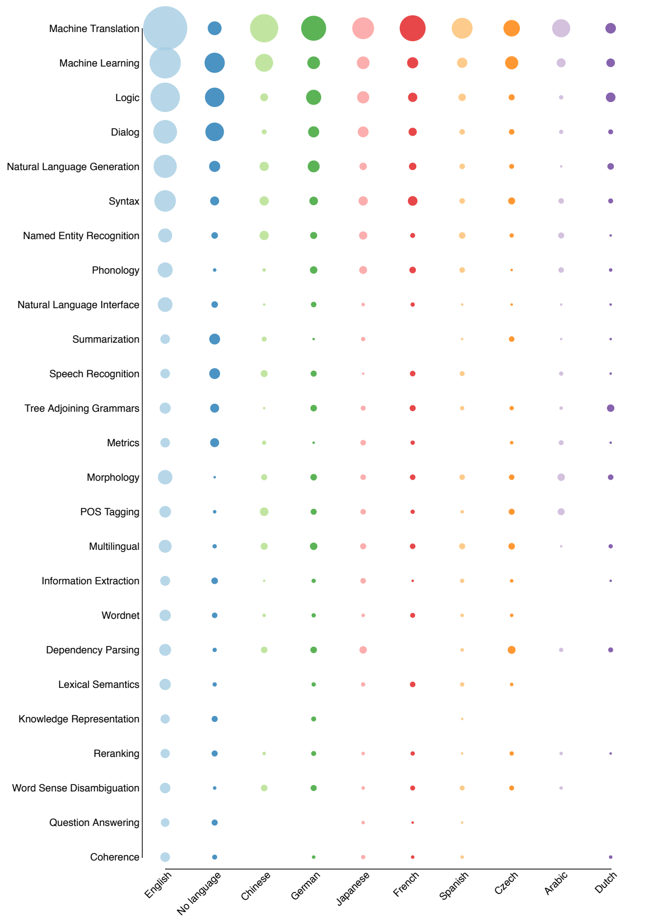

Figure 2: Categorical bubble chart that shows the relationship between research topics (y-axis) and the 10 most common languages in the corpus (x-axis). The bubble size encodes the number of papers.

4 Evaluation and Discussion

Since the visualizations were not experimentally evaluated, I will perform a task-based evaluation for each of the visualizations.

The first visualization allows users to perform several tasks: tracking the area of a language over time, comparing the width of the area of language over time (through the hovering interaction) or with another language (in the standard view). It encodes 3 data attributes simultaneously in an easy to grasp way.

A task that users might want to perform in this visualization is to get the exact number of papers for a specific language for a specific year. In the current visualization, the numbering on both the x-axis and y-axis is very rough and it would be insufficient to get an exact number. This was a design decision, to focus more on the rough pattern, but it is important to mention that it has downsides as well. Besides that, the standard problems with stacked area charts also exist here: especially around the peaks, it is hard to compare the width of the areas. This is something else that should be considered: it is more important to have an overview of the whole, or to be able to trace each area precisely?

The second visualization also supports several tasks: comparing frequency of research topics, comparing languages and - more importantly - tracking how frequent certain language-research topic combinations are. The added tooltip allows for more precision when comparing bubble sizes. This type of visualization in a grid makes it easy to spot patterns that stand out.

A task that is not supported here is sorting of the data. In this kind of grid, users might want to reorder rows and columns so that they can see clusters come together. This is something that is currently not supported, but would be a great feature to add in future work.

Overall, these visualizations are a first step in visualizing language diversity in the ACL Anthology Reference Corpus. They have shown that visualizing data that is not directly apparent from the corpus data can give us helpful insights about the current state of the field. Visualizations like these are also instrumental in raising awareness in the NLP research community, both about under-representation of languages and of assuming English as the "default" language.

References

[1] A. Anderson, D. McFarland, and D. Jurafsky. Towards a computational history of the acl: 1980-2008. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, ACL '12, pp. 13–21. Association for Computational Linguistics, Stroudsburg, PA, USA, 2012.

[2] E. M. Bender. The# benderrule: On naming the languages we study and why it matters, 2019.

[3] S. Bird, R. Dale, B. J. Dorr, B. Gibson, M. T. Joseph, M.-Y. Kan, D. Lee, B. Powley, D. R. Radev, and Y. F. Tan. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. 2008.

[4] J. Heer, M. Bostock, V. Ogievetsky, et al. A tour through the visualization zoo. *Commun. Acm*, 53(6):59–67, 2010.

[5] C. Kirov, R. Cotterell, J. Sylak-Glassman, G. Walther, E. Vylomova, P. Xia, M. Faruqui, S. Mielke, A. D. McCarthy, S. Kübler, et al. Unimorph 2.0: Universal morphology. *arXiv preprint arXiv:1810.11101*, 2018.

[6] J. Nivre, M.-C. De Marneffe, F. Ginter, Y. Goldberg, J. Hajic, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, et al. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1659–1666, 2016.

[7] B. Weitz and U. Schäfer. A graphical citation browser for the acl anthology. In *LREC*, pp. 1718–1722, 2012.