Digital History with Python

Analysing bibliographic records with PANDAS and SPACY

Mentor: Laura Fernández Gallardo

Mentee: Giuditta Parolini

Like data science, historical research is question-driven.

Why did the Roman Empire fall?

Is women's underrepresentation in computing historically motivated?

What are the roots of Big Science?

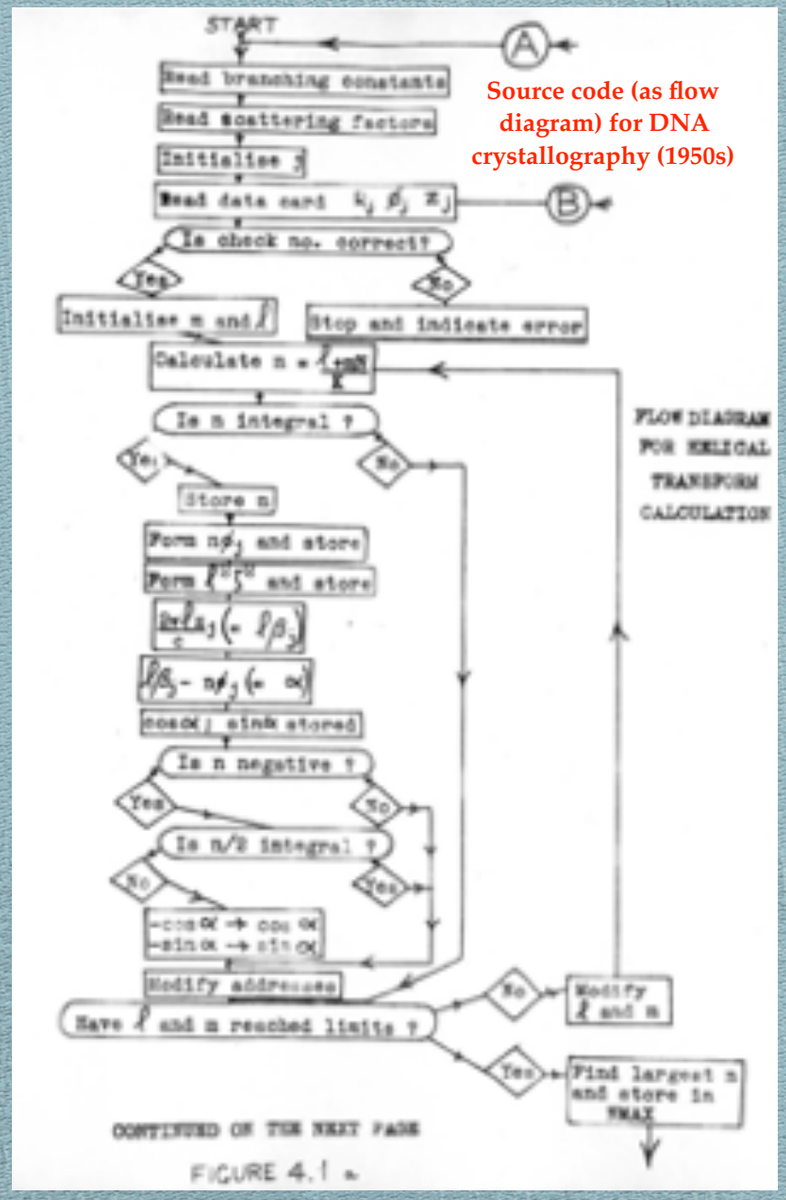
Did technology impact on social change in the Middle Ages?

Is the history of ICT an evolutionary or a revolutionary history?

Tell me your question and your primary sources*…

*primary sources (i.e. historians' data) are letters, diaries, newspaper accounts, books, photographs, census statistics, etc. dating back to the time of the events

… and I will tell you what kind of historical account you are going to write.



Source code (as flow diagram) for DNA crystallography (1950s)

History must be global, not focused only on the Western World	History must take into account women, minorities, and other underrepresented groups	There is more to history than kings, entrepreneurs, and Nobel Prize winners. E.g. there are artifacts.

Today

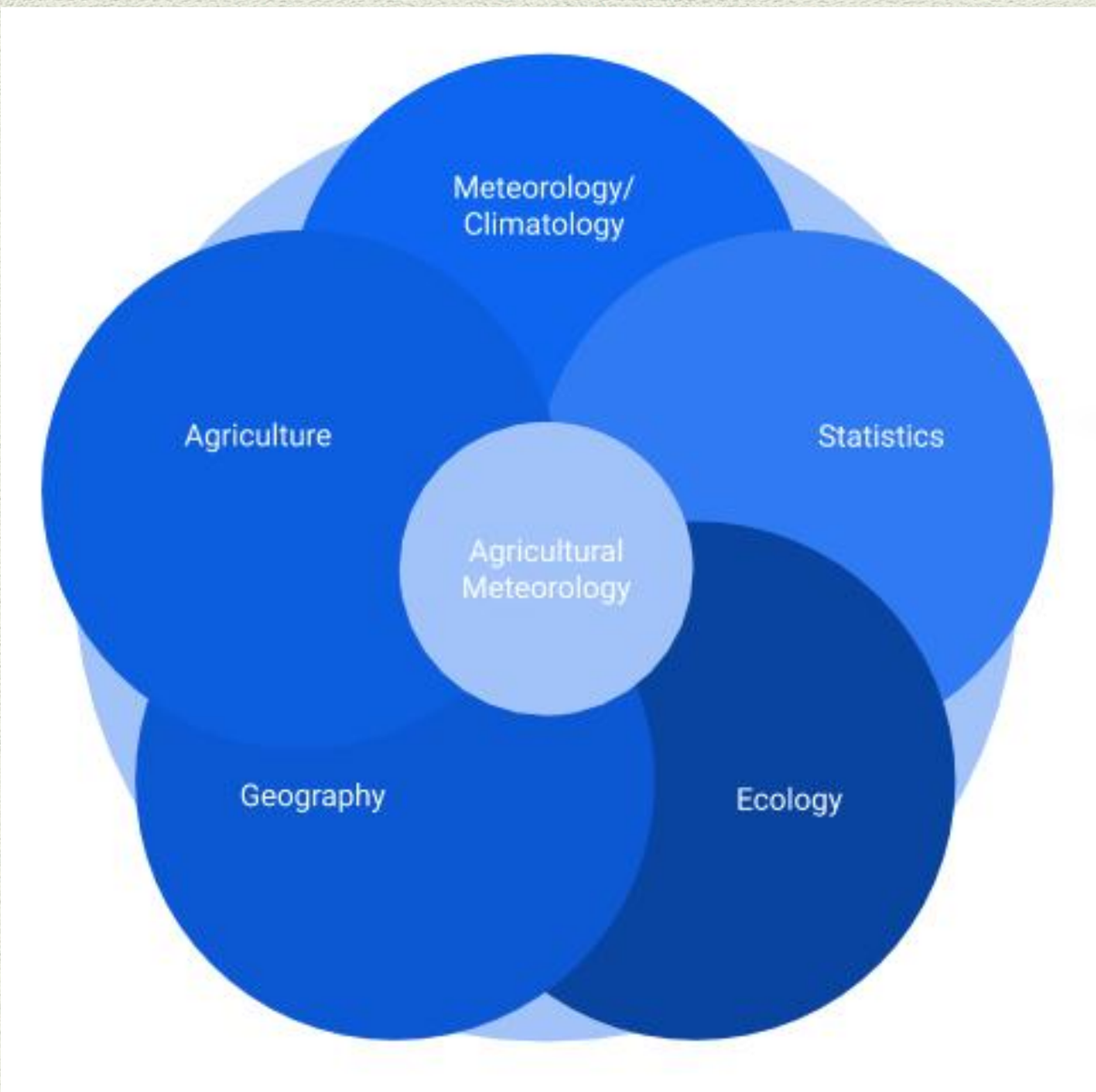
Historians have broad views about the questions worth asking

Can data science contribute to broaden further historians' research questions and the sources they can use?

It is not the farmer, but the good weather that makes the corn grow (Fuller 1817)

My Project

History of agricultural meteorology (1900-1950)



AM is an interdisciplinary research field that studies how weather and climate affect crop growth and livestock performance.

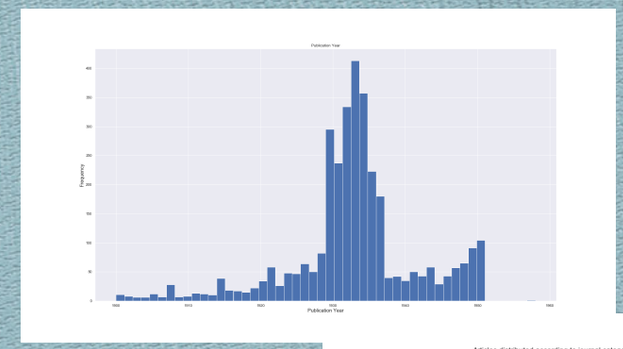
In the first half of the 20th century, international organisations and national weather and agricultural services began to promote systematic research in agricultural meteorology

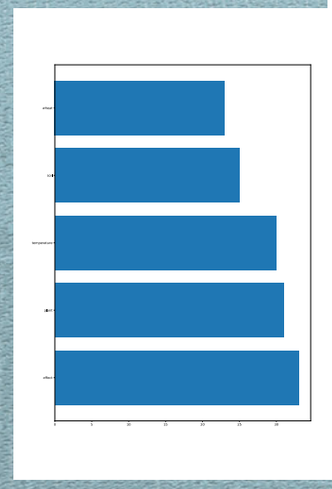
https://agriculturalmeteorology.wordpress.com

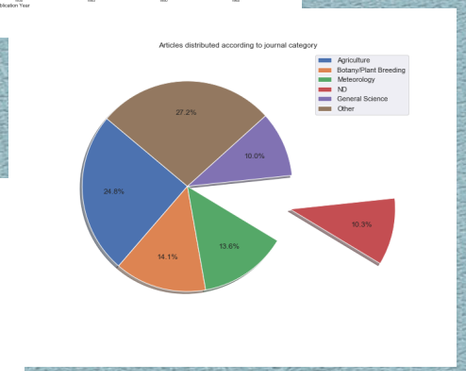
Questions

- **Can we investigate the interdisciplinary nature of agricultural meteorology from its publications?**
- **Who were the authors of these publications?**
- **In which field did they do their research?**
- **In which journals did they publish their papers?**
- **What were the most popular topics in agricultural meteorology?**
- **Did the number of publications increase over time?**

WORK PROCESS







Zotero library (exported as csv file). Full-text of a few articles extracted as txt for analysis

Data cleaning and Python coding using Pandas and SpaCy

Data analysis and visualisations

Data and data cleaning

- Data set: Bibliography of Agricultural Meteorology (1900-1950) collected as part of my academic research project.
- The bibliography includes books, journal articles, thesis, reports on agricultural meteorology printed between 1900 and 1950 in several languages.
- The data are stored in a Zotero library (https://www.zotero.org).
- The data set is a work in progress (so far over 3.300 entries). The original data have to be checked and standardised. Information on languages and journal full titles is still incomplete.
- The data cleaning is done manually at present. There are known issues with non-English characters and typos added at the data entry stage.

Correspondence Journals - Categories - Sheet1

Publication Title	Category
Acoreana	Natural History
Acta Agralia Fennica	Agriculture
Acta Jutlandica	General Science
Acta Phaenologica	Agriculture
Acta Societatis Botanicorum Poloniae	Botany/Plant Breeding
Actes de la Société linnéenne de Bordeaux	Natural History
Actes et Comptes Rendus de l'Association Colonies-Sciences	Agriculture
Agrario levantino	Agriculture
Agricoltura coloniale	Agriculture
Agricultura Sinica	Agriculture

Interdisciplinarity

Each journal is associated (by hand) to a category. In total, there are over twenty categories ranging from 'Agriculture' to 'Hydrology'.

Part 1 / Using Pandas

- Inspecting the data (e.g. checking for missing values).
- Extracting valuable information on authors and publications.
- Visualising the data set.

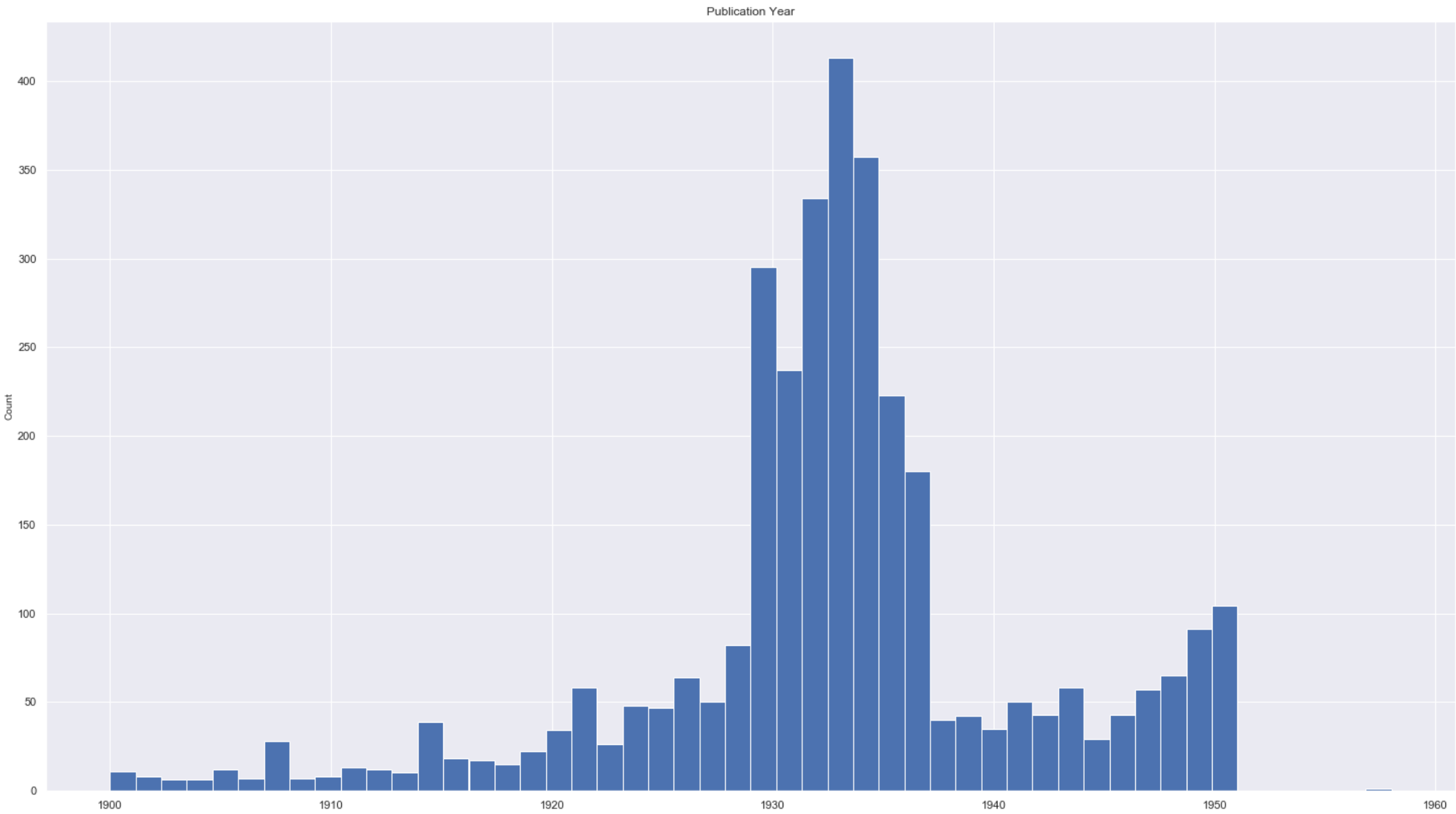
Part 2/Using SpaCy

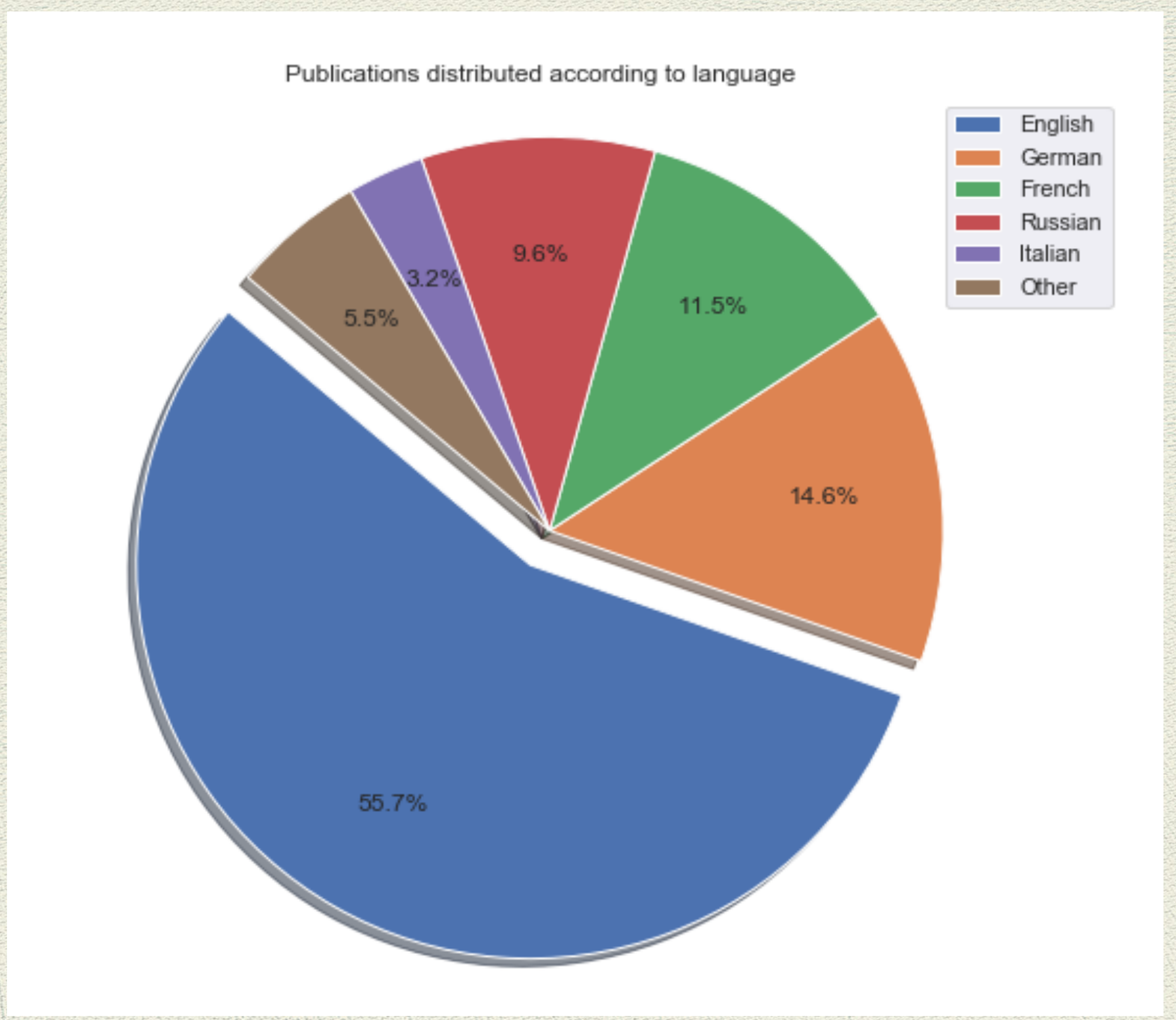
- Full-text analysis of a small sample of journal articles listed in the bibliography. The sample is small because the manual extraction of the text is time consuming. The pdf files are obtained scanning the original paper copy. OCR software is used for character recognition, but the results are poor.

- Text analysis of all the titles of English journal articles listed in the bibliography.

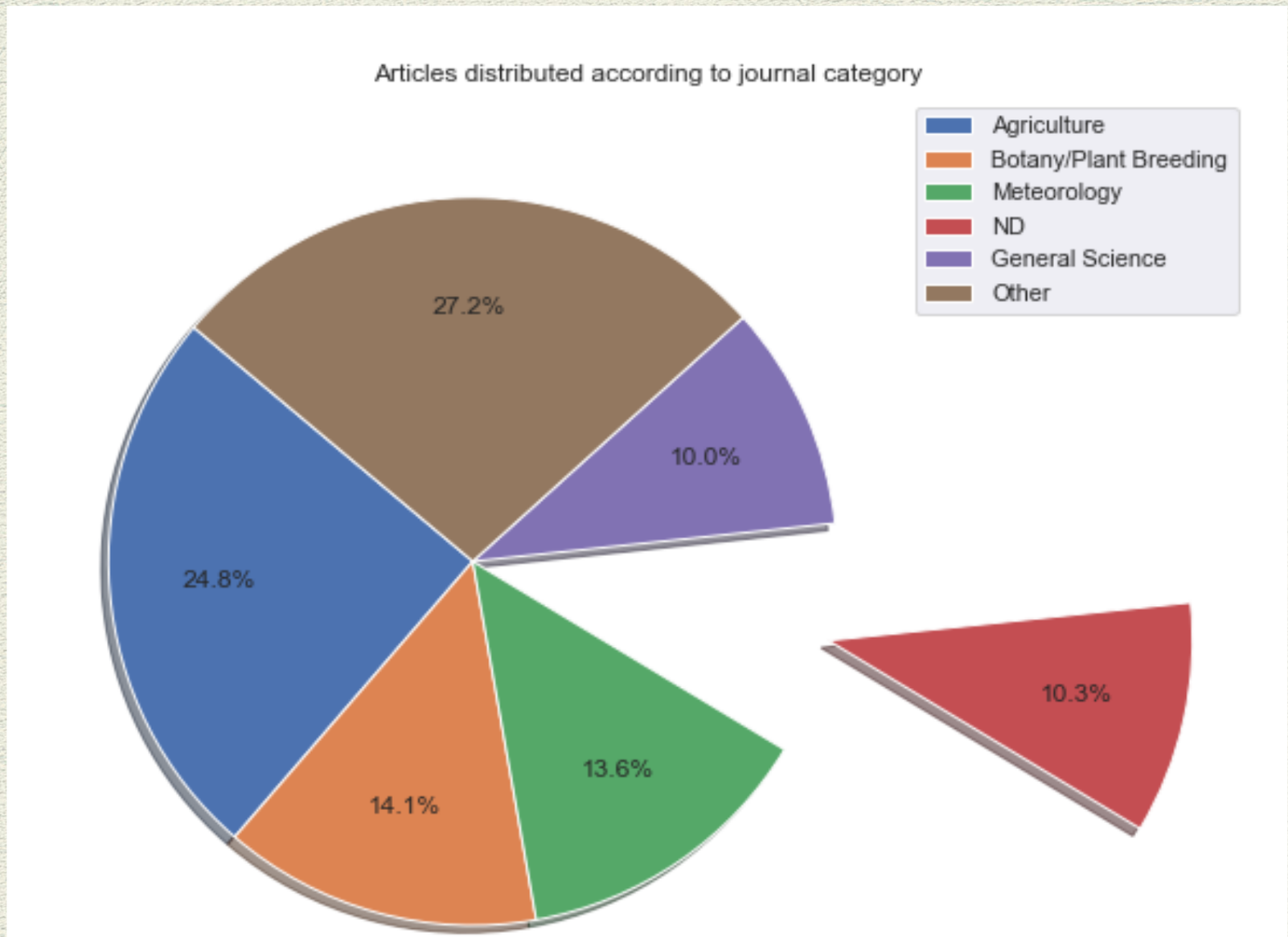
Part 1 / Preliminary results

The number of publications per year has a peak in the early 1930s. This is just an artifact of my data set, because several entries have been extracted from bibliographies of papers published in that period.



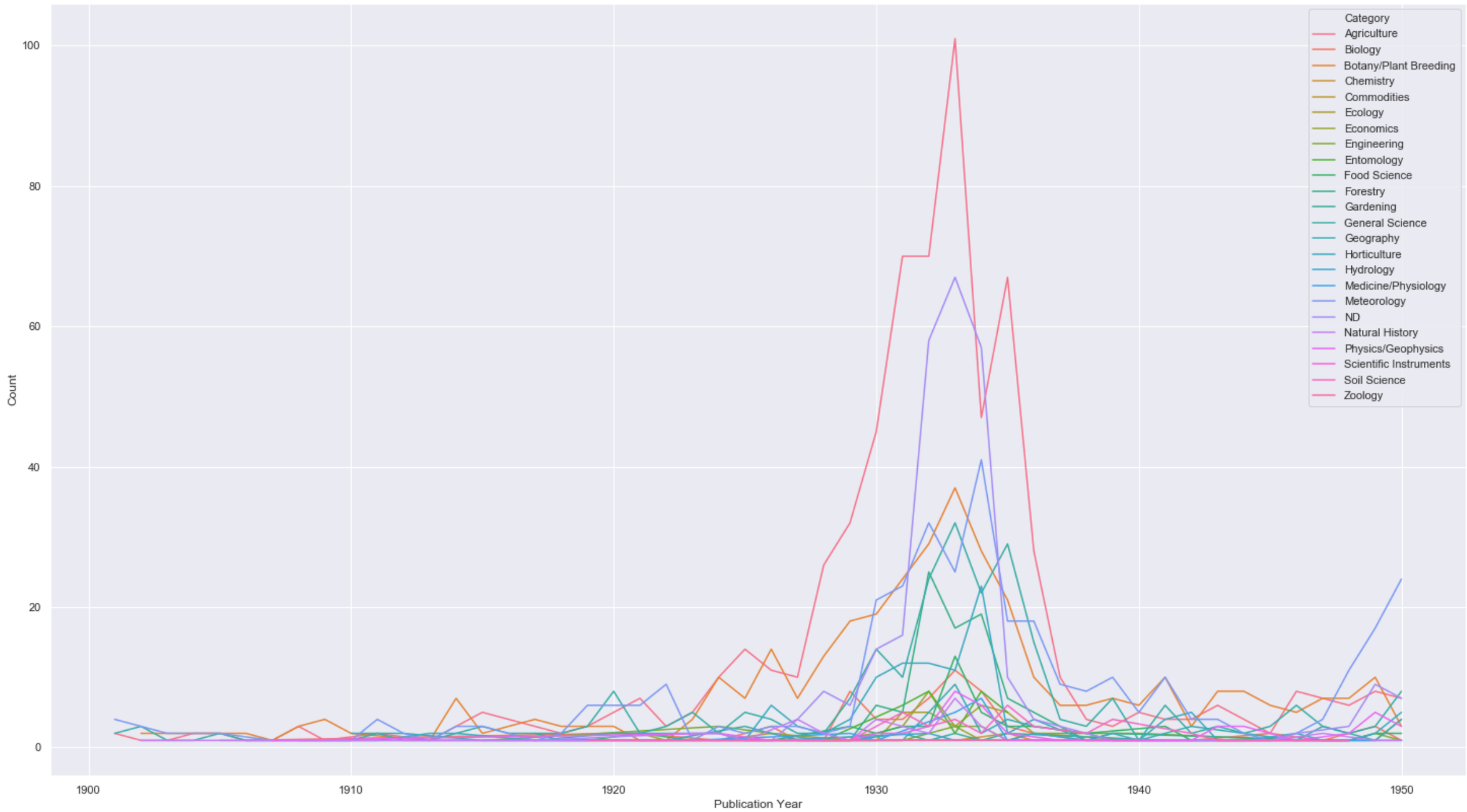


There are publications in over twenty languages in the data set. English is the main language (55,7%), but over forty per cent of the references are to publications in other languages. This is not surprising because in the first half of the twentieth century, French, German and other national languages were also commonly used in science.



Almost 25% of articles were published in agricultural journals, and less than 14% in meteorology journals. This may -or may not- be an artifact. Several bibliographies used to acquire data were compiled by agricultural departments. About 10% of the publications cannot be assigned to any category (ND), because there is not enough information on the original journals.

The number of publications per category has a peak in the early 1930s. This is an artifact of the data set, due to the increased amount of publications available for that period. In the late 1940s, the amount of meteorology publications starts to increase. As the data set ends in 1950, it is not possible to say whether this increase is significant or not.



Authorship

- The list of authors contains over 2.600 names. But only 14 of these people authored or co-authored ten or more publications.

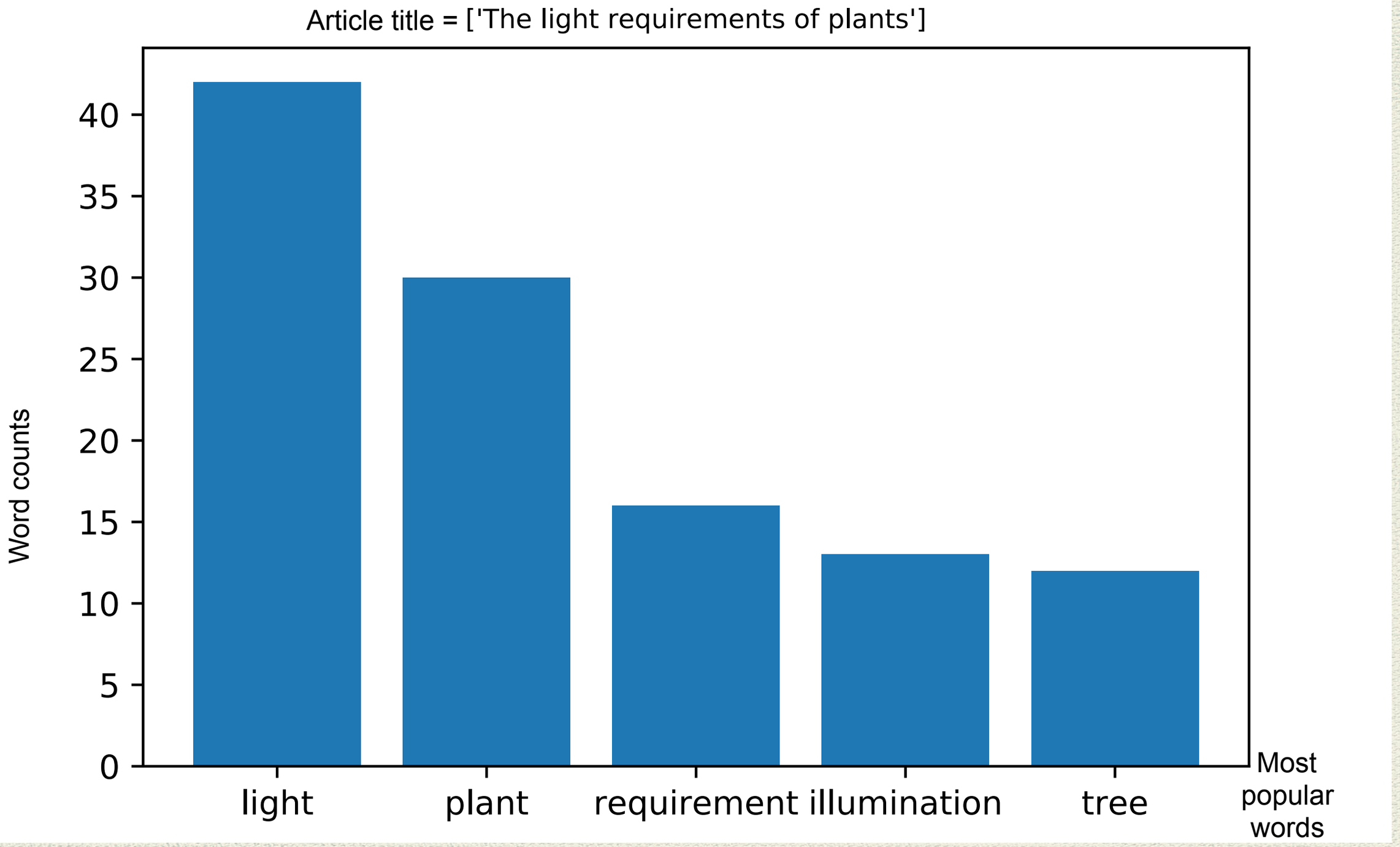
- The articles appeared in almost 800 journals, but only 7 of these journals (0,9%) published 50 or more papers.

Co-authorship

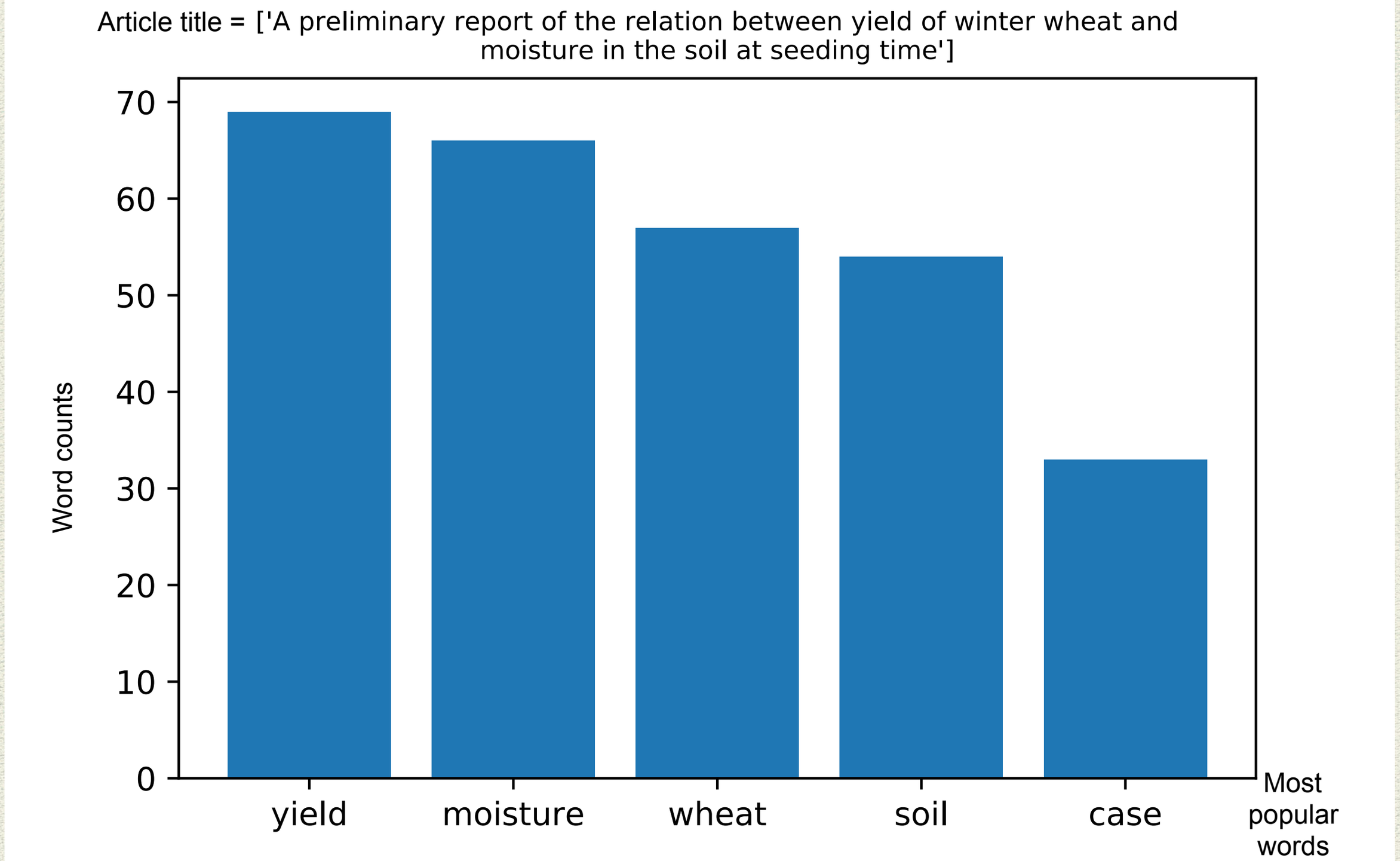
With my code, I can extract the list of people who coauthored papers. The co-authors' data can be used for network analysis.

Coauthor(s) of Aamodt, O. S.: Platt, A.W.
Coauthor(s) of Afzal, M.: Throught, T.
Coauthor(s) of Aikins, G. A.: Fay, A. C.
Coauthor(s) of Aikman, J. M.: Dodge, A. F.
Coauthor(s) of Akamine, E.: Clements, H. F., Moriguchi, S., Shigeura, G.
…
Coauthor(s) of Zscheile, F. P.: Beadle, B. W., Roach, J. R., White, J. W.
Coauthor(s) of Zvorykina, K. V.: Elagin, I. N.

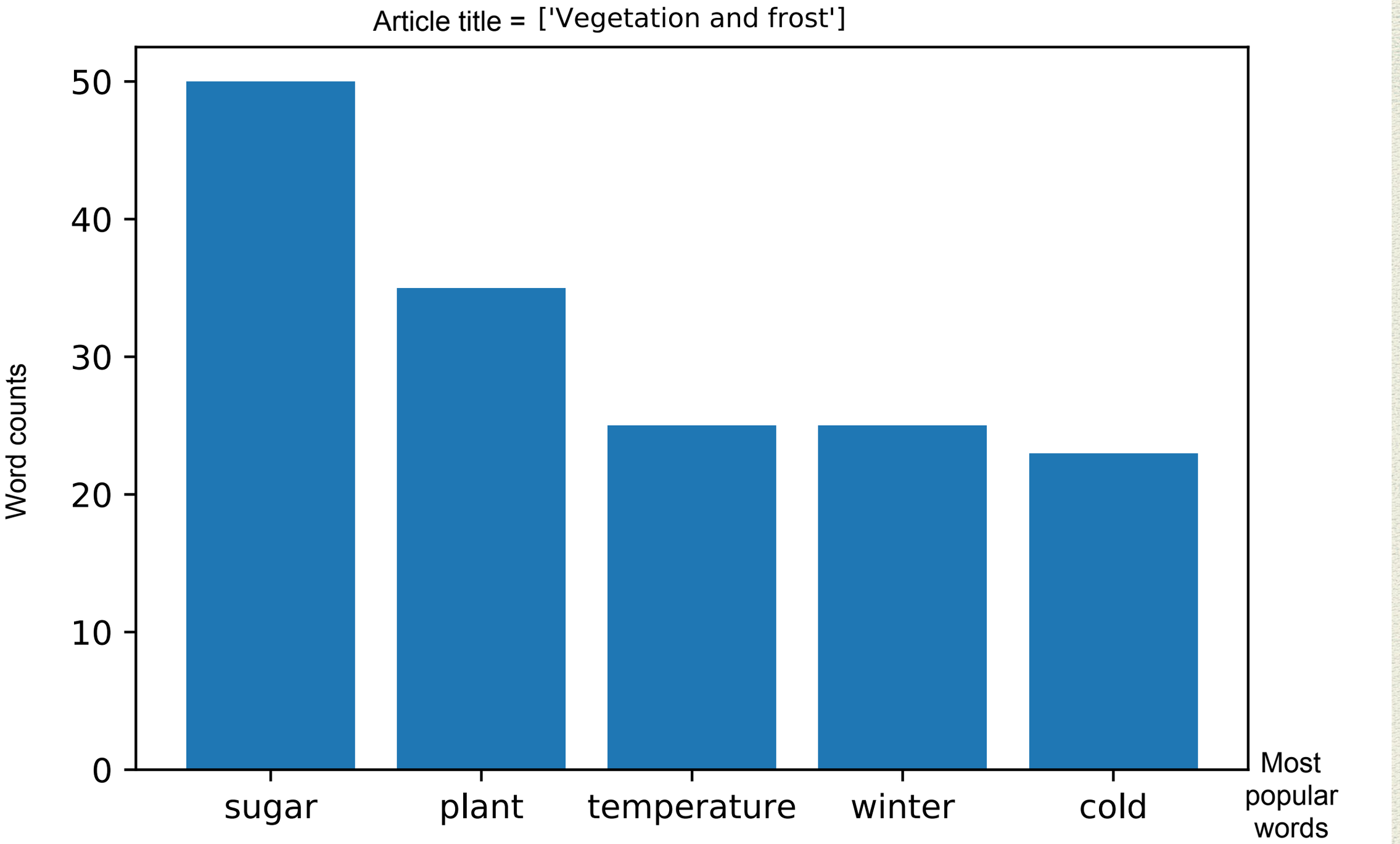
Part 2 / Preliminary results



For each of the full-text articles, the code extracts lemmas, removes stop words and generates plots of the most common words in each text, as in this histogram.



This plot displays the five most popular words in the full-text of an article on winter wheat. In this paper, four recurrent words out of five are also included in the title, but title words are not always keywords.



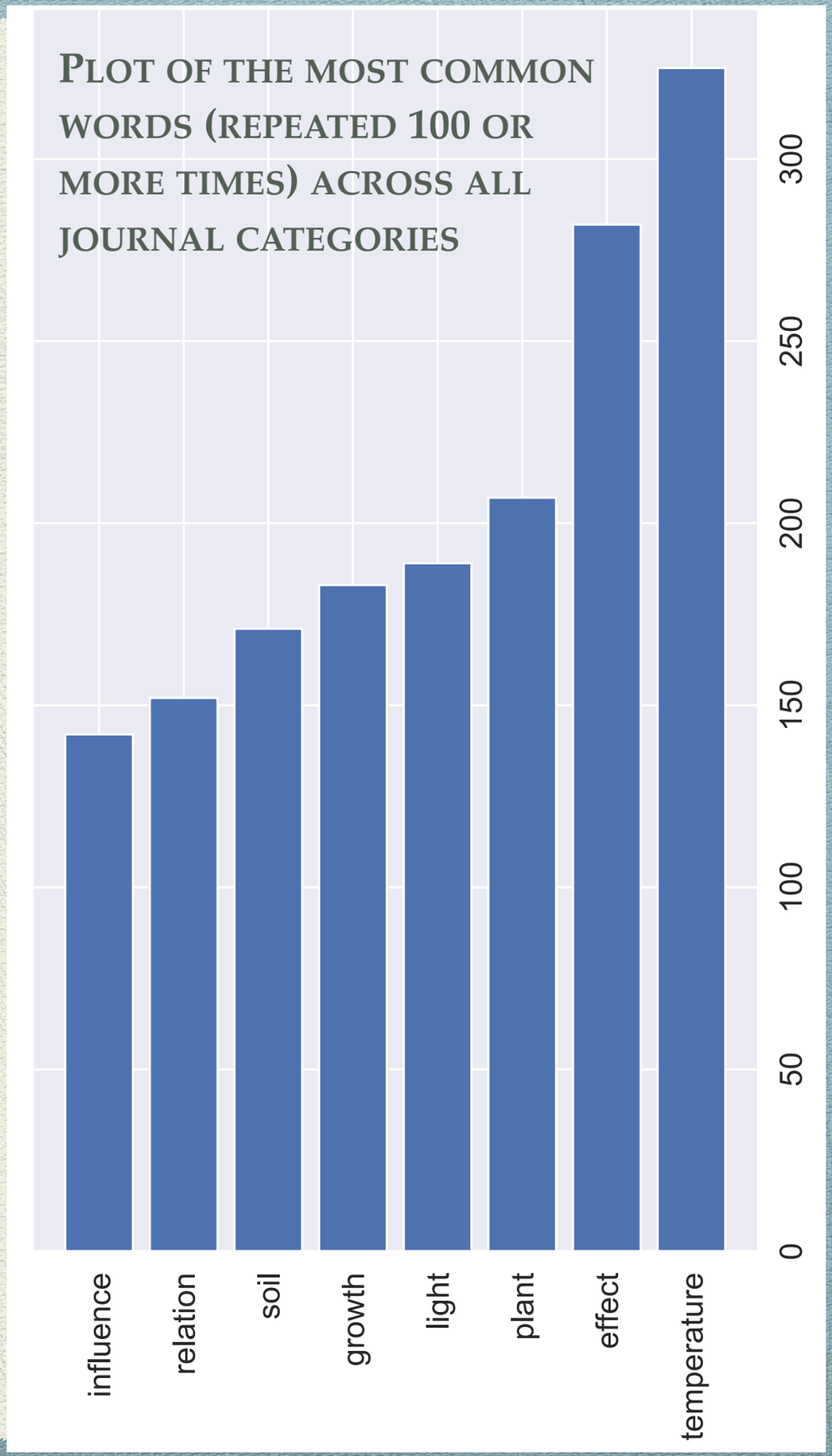
In this other paper on frost damage, the title is very short and neither of the words is a recurrent lemma in the full-text. However, recurrent words in the text are semantically related to the words in the title. E.g. plant-vegetation; winter-frost.

Where did research in agricultural meteorology take place?

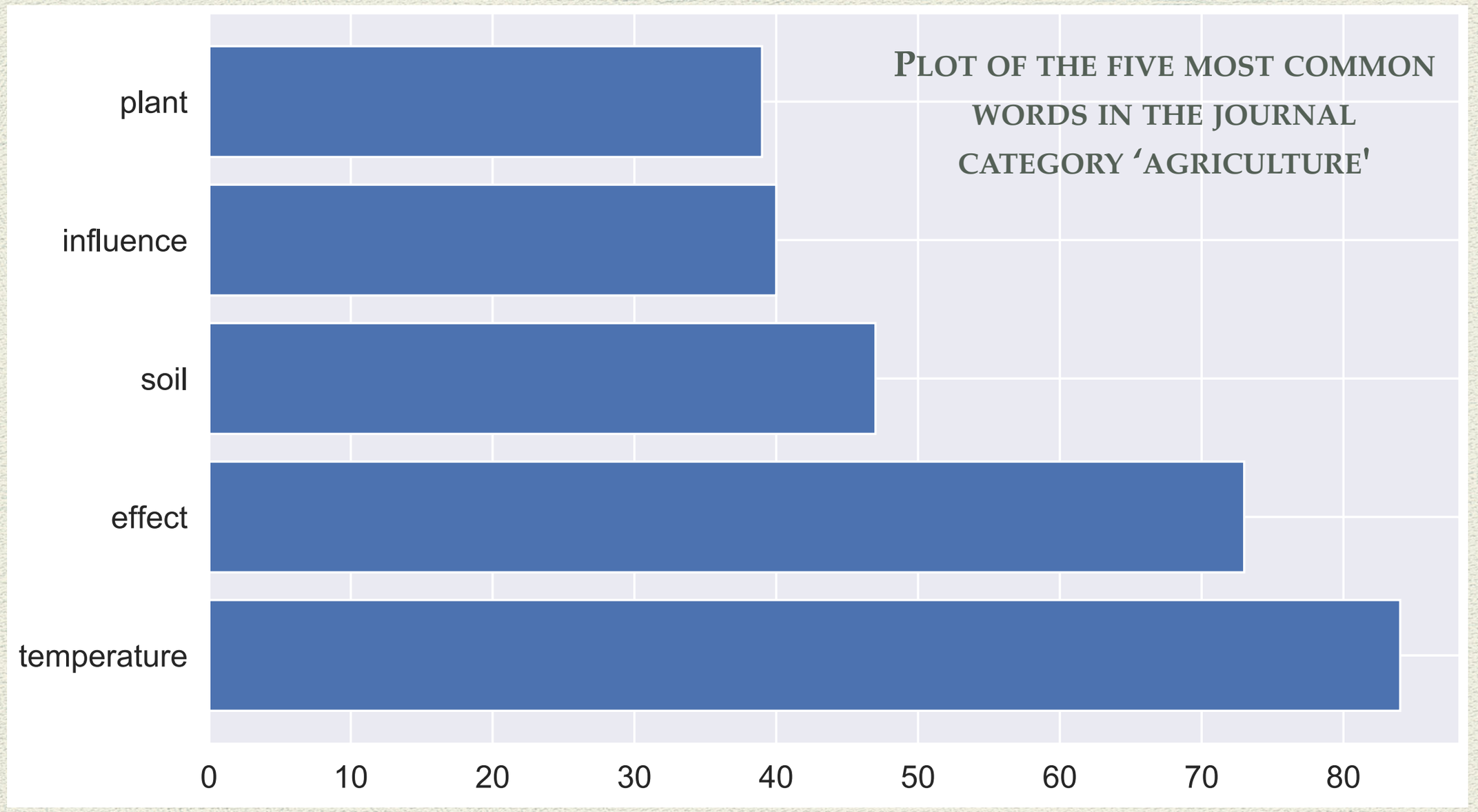
- Testing named entity recognition (NER) for geographical names with SpaCy. The full-text of the English journal articles available is examined.
- Results are not satisfactory. The model cannot always discriminate between geographical names and other entities (scientific names of plants, names of scientific institutions including a geographical entity [e.g. US Weather Bureau], etc.).

Title analysis of the English journal articles

The most common word is 'temperature'. The word 'plant' comes only third in the list. Evidently, temperature was considered the key variable in agro-meteorological studies. Popular words were also 'effect', 'relation', 'influence'. This suggests that many agro-meteorological studies were correlation studies in this period.



Title analysis of the English journal articles



Agriculture is the category for which there are more journal articles available (tot. 663). Also in this category, 'temperature' is the most common word. Surprisingly, the word 'plant' comes only fifth in the list.

Do you remember my questions?

- **Can we investigate the interdisciplinary nature of agricultural meteorology from its publications?**
- **Who were the authors of these publications?**
- **In which field did they do their research?**
- **In which journals did they publish their papers?**
- **What were the most popular topics in agricultural meteorology?**
- **Did the number of publications increase over time?**

Do they have answers?

Yes, they do.

These answers are provisional because the data set is neither complete nor adequately cleaned, but some features are already evident…

In the first half of the 20th century agricultural meteorology was:

- A (geographically) widespread field of research with publications produced across the globe in a multiplicity of languages.

- A fragmented field with contributions written by thousands of authors and printed in hundreds of different journals belonging to over twenty different categories.

- A research field, whose keywords were ‘temperature’, ‘plant’, ‘light’, ‘growth’, ‘soil’, and whose investigations mainly concerned relations between weather / microclimatological factors and plant growth.

Acknowledgments

I sincerely thank my mentor, Dr Laura Fernández Gallardo, for the support received during the Learn IT, Girl! initiative.

The bibliographic dataset is being built with funding provided by the Deutsche Forschungsgemeinschaft (DFG) (Project No. 321660352).