ΔΠΜΣ Γλωσσική Τεχνολογία

Thesis Outline

Computational Analysis of Scientific Discourse.

a) https://explore.openaire.eu/fields-of-science#01%20natural%20sciences

b) https://sci-hub.wf/

Certainty and uncertainty in the English Language is expressed in various ways. From modal auxiliaries (*may, might, could, will, must, can/can't, would*) and quasi modals (*be bound to, have to*) to modal adverbs (*certainly, of course, undeniably, probably, allegedly, perhaps*), modal adjectival and participial constructions (*likely, uncertain, doubtful*, *convinced, certain*), modal nouns and even lexical verbs (*assume, believe, think, conclude, imply, infer*). These markers **of epistemic modality** could help us gain insights and draw conclusions about the use of language in different scientific communities, how each scientific community expresses their scientific claims -the authors being confident (or not) towards their findings, and finally detect and show any possible relationship/correlation (positive or negative) between certainty in scientific discourse and article impact/visibility. #Biotechnology vs Medicine, Climate, AI, Physics | Temporal dimension

#Find retracted papers dataset

This study seeks answers to the following questions:

1) In what ways does each scientific community (physics, mathematics, social sciences, biology, engineering, computer science, medicine) express their scientific claims, as far as certainty is concerned?!

2) Do certain epistemic modality markers tend to cluster in particular scientific domains?!
3) In what ways are they distributed across different scientific domains?!
4) Does a very confident discourse/language choice affect the visibility of the paper?! ***#Reproducibility***
5) Is discourse certainty/degree of confidence towards research claims/accomplishments a marker of a paper's **positive** course over the years?! (number of citations etc.)
6) Is discourse certainty/degree of confidence towards research findings/accomplishments a marker of a paper's **negative** course over the years?!
7) Do less confidently expressed claims get the attention they deserve?! ***#More of a conclusion***

Key-words: **certainty, epistemic modality**, article influence, article visibility, abstracts, patterns, authorial stance, discourse markers, author certainty, **article citation resilience**, scientific discourse analysis, science genre, rhetorical structure, modality.

Key-tasks: Find the **language markers** that express certainty and "big" language. Decide the tools that you are going to use. SciARK and ChatGPT or GPT-4 (you will have to pay).

Draft-Outline

1. Gather scientific articles from big data pools.
2. Classify them to the aforementioned categories. Notes: **a)** How many categories am I going to analyze?! **b)** How am I to classify the articles?!
3. Infer the articles of each category and get the conclusion sentences, that is the claim sentences. Experiment with different tools, like SciARK, ChatGPT and ChatGPTplus (GPT-4) -for this you will need an API and a good prompt.

4. Analyze the claim sentences. Create lists with words for Certainty and Uncertainty. It's not necessary to get too binary, we can let loose and have a spectrum that begins with pure certainty and ends with pure uncertainty. Notes: **a)** Count just words? Or use n-grams?
5. Check the articles of each category and see what words of certainty or uncertainty are predominant. Report the results.
6. Now the fun part! For each article of every category, see if the articles with the most expressed certainty towards their research claims, have a good visibility/reproducibility course, over the years.

Disclaimer

My analysis will be a linguistic/discourse one. Therefore, the study is not going to check if the claims (independently of their author's certainty towards them) are true or not. So, for example, if words that express big certainty are proven to be used in articles in social sciences, and these articles (with the high certainty markers) are also cited the most by this community -that has nothing to do with the truthfulness of the claims per se. It will be considered just an indicator that language has the power to shift our attention to sentences and ideas that are powerfully expressed, independently of their truthfulness.