Universiteit Leiden

Questions and disinformation: a cross-linguistic comparison

Applicants

Eligible proposals must have at least two applicants from Humanities, preferably with an interdisciplinary approach.

Supervisor Name	Discipline
Matthijs Westera	Assistant professor Humanities & AI
Jenny Doetjes	Professor Semantics & Language variation

Project description

Provide a brief description of the project (max. 300 words)

In a previous project of the applicants, two multilingual datasets of tweets have been established; one containing tweets with hashtags related to the COVID-19-pandemic that strongly correlate with **disinformation** (from now on, **COV#D**) and one with tweets containing **neutral** hashtags (**COV#N**). Each dataset contains data from four different languages: English, French, Italian (10.000.000 tweets per language) and Dutch (3.000.000 tweets). The utterances in the tweets were automatically tagged as non-questions and (different types of) questions. In addition, they were classified based on the occurrence of features that are associated with disinformation in the literature.

Given the fact that the relevant literature mostly considers declarative sentences in English (that is, 'ordinary sentences' or statements), the created multilingual datasets offer a unique opportunity to investigate to what extent the generalizations reported in the literature also apply to questions, and whether similar generalizations hold for different languages. A first, superficial investigation of the datasets strongly suggests that there are both important differences between questions and non-questions, and between the four languages in the datasets.

The current project aims at an in-depth investigation of the most striking differences between the datasets. Take for example negation, a feature that has been associated with disinformation. This is confirmed by the datasets for English declaratives, but, unexpectedly, not for English questions, which turn out to contain less negations in COV#D as compared to COV#N. For the Dutch data the pattern was the other way around: while the declaratives had less negations in COV#D, the questions had more negations in COV#D. The goal of the project is to analyse this and other differences between the two datasets, in order to gain a better understanding of linguistic features associated with disinformation in various languages and of the way these features interact with sentence type (questions vs non-questions).

Research Trainee Profile

Each proposal requests two Research Trainees. Describe the general tasks of the research trainees, how these tasks are academically challenging to the research trainees, whether they need any preliminary knowledge (regarding the topic and/or research methods) and which skills the research trainees should have. Also specify which type of students are eligible to apply (3rd year Ba, Ma, ResMa).

We are looking for Ma or ResMa students with a background in Linguistics, Natural Language Processing, or Journalism. Candidates should be fluent in at least two of the four languages in the dataset (English, French, Italian and Dutch).

The following properties are a plus, but not strict prerequisites:

- Demonstrable courses or term papers on relevant topics, especially pragmatics, semantics, questions, rhetoric, or disinformation.
- Demonstrable familiarity with relevant research methods, such as linguistic annotation, corpus analysis and natural language processing (e.g., Python).

Collaboration

If applicable: Describe how your research improves collaboration and cross-pollination between the disciplines involved (max. 300 words)

The proposed traineeships will contribute to the interdisciplinary collaboration between the applicants on questions and disinformation. Matthijs Westera's ongoing research focuses on the role of explicit and implicit questions in discourse structure, relying on quantitative methods, big data and machine learning. Jenny Doetjes was one of the applicants of the recently finished NWO-project *Understanding questions,* which examined the relation between prosody, processing, syntax and semantics in French and Mandarin by means of theoretically driven experiments. Last year, they started a data-driven project on questions and disinformation (via the traineeship program). The proposed traineeships aim at providing a more fine-grained investigation of the datasets that were created during the previous project.

Our collaboration and the proposed traineeships will also increase synergy between groups within LUCL that are not normally working together. Given the topic of our project and its combination of methodologies, the results will be relevant for both theoretical and computational linguists, as well as researchers studying disinformation from the perspective of language use and journalism. Given the focus of the project on cross-linguistic variation, the results will also be relevant for colleagues in LUCL working on linguistic diversity and comparative linguistics.

Deliverables

Enumerate intended project results: papers, research proposals or otherwise. (max 200 words)

The project is intended to result in the following:

- A thorough analysis of the COV#D dataset (with COV#N as a control), focusing on cross-linguistic variation of features that are associated with disinformation and their interaction with sentence type (questions vs. declaratives).
- One academic paper reporting on the foregoing items as well as the resulting analysis, submitted to a fitting conference such as the International Pragmatics Conference, Semantics and Pragmatics of Dialogue (SemDial), Linguistic Resources and Evaluation (LREC), various computational venues (ACL, EMNLP, CoLing, IWCS), more theoretical venues (SALT, Sinn und Bedeutung) or one of the many student research conferences.

Planning

Provide a breakdown of the project into phases with tentative timing (max 150 words)

Month	Activities	Hours*
1	**Warming up.** Initial exploration of literature on features associated with disinformation and the datasets.	25
2	**Stage 1: Selection of the phenomena for the in-depth cross-linguistic investigation.** Evaluation of the quantitative data resulting from the previous project in view of the literature on features associated with disinformation.	25
3 - 6	**Stage 2: Qualitative and quantitative analysis.** The trainees will develop and pursue their own hypotheses and evaluation methods, depending on their interests and skills, focusing on differences between at least two languages within the dataset.	100
7	**Finalize paper.** During each stage the trainees will already work on a first draft of a corresponding section for the paper, which will be completed and submitted in the final month.	25

*Approximate working hours *per trainee*, based on 7 months × 0.15fte = 173 hours per trainee.

Student Application

Provide information on how to apply e.g. required documents for application (resume, motivation letter etc.) and an email address where student applications should be sent to.

Please submit your application, consisting of a resume and cover letter in a single PDF file, to m.westera@hum.leidenuniv.nl.