Corpus Pragmatics (2026) 10:19
https://doi.org/10.1007/s41701-025-00206-4

RESEARCH

Constructing Expertise Through Questions?—Potentials of Corpus Pragmatic Regression Analyses

Gunilla Kaibel[1]

Received: 18 June 2025 / Accepted: 10 September 2025

Abstract
This study challenges traditional assumptions about questioning practices by examining how questions, conventionally associated with knowledge gaps, can surprisingly serve to construct expertise. Using corpus-pragmatic regression analyses, the research investigates how different types of questions serve as tools for taking epistemic stance in online scientific discourse among various groups of discourse participants. The analysis reveals statistically significant differences in linguistic behaviour not only between experts and laypeople but also within the lay community. I develop an expansion to existing models of epistemic hierarchies in questioning to account for more possible epistemic hierarchies between questioner and addressee.

Keywords Quantitative linguistics · Epistemicity · Questions · Expertise · Expert-laypeople-interaction · Non-canonical questions

Introduction

Questions are generally presumed to express a desire to elicit information and thereby close knowledge gaps. Consequently, questions could be interpreted as signs of missing expertise. However, questions can fulfil diverse discursive functions beyond information seeking, e.g. expressing doubt or initiate repair (cf. Stivers, 2010: 2776). This raises the question: how are different kinds of questions used as a tool to enact expertise instead of communicating lack of knowledge?

This paper is the first to quantitatively investigate how questions are employed by different discourse participants to construct expertise. I examine questioning practices in comment sections of German scientific blogs over two decades. In these sections, experts actively engage in discussions with laypeople. Interestingly, the

✉ Gunilla Kaibel
Gunilla.Kaibel@ds.uzh.ch

[1] Deutsches Seminar, University of Zurich, Zurich, Switzerland

Published online: 23 December 2025

relative frequency of comments containing questions is similar across groups. But which questions do the different groups of discussants use and what functions do the questions fulfil? The goal of this paper is to determine how different question types correlate with different groups of discussants.

I hypothesise that specific types of questions to fulfil particular pragmatic functions and that the distribution of question types varies significantly among groups of discourse participants depending on their (assigned) expertise. The manner in which participants pose questions plays a crucial role in demonstrating their level of knowledge. Significant differences give insights into recurring commenting practices and epistemic stances of discussants.

Employing regression analysis for corpus pragmatic purposes, my approach combines theoretical insights about questions' epistemic and social functions with robust quantitative analysis. It advances quantitative methods in corpus pragmatics by applying regression analysis in a novel direction: it examines how linguistic patterns correlate with non-linguistic variables in order to identify pragmatic functions.

The first Section ("Theoretical Background") outlines the theoretical background including conceptualisations of experts and laypeople as well as different research approaches to questions and their functions. Thereafter, I describe the methodological framework (Section "Methodology and Methods"). After describing the corpus data (Section "Corpus Data"), I describe the classification process used for analysis (Section "Classifying the Data – Creating Variables"). Section "Combination of the Variables: Regression Analysis" presents the findings of the regression analyses while Section "Closer Look at Possible Functions: Annotating Vielleicht-Questions" presents the results of further annotation and connects them to concepts of politeness and face. "Discussion and Outlook: What do the Results Show About the Construction of Expertise?" examines how these results enhance the understanding of expertise construction in comments. The last Section "Conclusion" concludes this paper by summarizing how expertise manifests itself linguistically in discourse.

Theoretical Background

Conceptualisations of Experts and Laypeople

The concept of expertise can be approached from multiple theoretical perspectives. Bock and Antos (2019) identify four key characteristics of experts:

1. Knowledge ('Sachkompetenz'[1]),
2. Theoretical expertise ('Theoriekompetenz'),
3. Innovation competence and
4. A position that authorises their expert-status (institutional, intellectual, prominence in media) (cf. Bock & Antos, 2019: 61).

[1] All German scientific terms, quotes and examples are translated by the author with support of *DeepL* and marked with 'single quotation marks'.

They emphasise that expertise is domain-specific rather than general, typically linked to professional specialization (cf. Bock & Antos, 2019: 60).

Spitzmüller (2021) focuses on the relational aspect of expertise, arguing that experts emerge through recognition by peers and laypeople. Expertise is established when people present themselves as experts or are presented as experts. This is done through alignment and differentiation from other experts and the respective laypeople (cf. Spitzmüller, 2021: 13–14). The role as experts gives them power in the form of *voice* and *agency* in discourse (cf. Spitzmüller, 2021: 4–5, 9).

The anthropologist Carr offers a compelling perspective by reconceptualizing expertise as performance rather than possession. She proposes that expertise is "something one does" (Carr, 2010: 26). Expertise is created through words, habitus, and discursive practices, which Carr calls "enactments of expertise" (Carr, 2010: 17). Language plays an important role in this process. She describes her "understanding of expertise as accomplished, or enacted, through linguistic patterns" (Carr, 2010: 19).

> After all, to be an expert is not only to be authorized by an institutional domain of knowledge or to make determinations about what is true, valid, or valuable within that domain; expertise is also the ability to 'finesse reality and animate evidence through master of verbal performance' (Matoesian 1990: 518). (Carr, 2010: 19)

Breeze (2023) differentiates between explicit and implicit strategies to ground authority. Explicit grounding includes mentioning reasons for superior knowledge. Implicit grounding is more subtle, e.g. by using special terminologies (cf. Breeze, 2023: 39–44), certain linguistic patterns (e.g. *ich als Experte* 'I, being an expert' (cf. Merten, 2024: 305)), a certain register of language like specialised language (cf. Simon & Janich, 2021: 44), or potentially also lexico-grammatical features such as specific types of questions which will be the focus of this study.

Just like *expert*, *layperson* is a relative category. One can be an expert on one topic and a layperson concerning another topic or in relation to another person. The question is whether *layperson* can be solely defined by the absence of expert status. Spitzmüller (2021) argues that only non-experts interested in the topic can be considered laypeople. What differentiates them from the experts is primarily their social position, e.g. that they possess no qualification like an academic degree, no institutional authority, and gain no economic reward for their knowledge (Spitzmüller, 2021: 4–5).

There are only a few linguistic studies about laypeople's discourses. Hoffmeister et al. 2021 collect several studies about what laypeople know and think about language and linguistic topics and how they present and discuss this knowledge. Much of linguistic literature focuses on lay discourse about language itself (e.g. Antos et al., 2019; Schrader-Kniffki, 2014). This fascination of linguists in laypeople's discussions about language leads to a research bias that discourse about topics other than language is considerably less studied.

However, there are linguistic studies on lay discussions about other topics: Merten, 2024 studies how laypeople discuss health-related topics in online comments, Coen et al. 2021 study expertise construction in comments about climate change.

The former studies usually do not discuss that there are different kinds of laypeople with presumably different communicative practices. Through a large corpus and quantitative methods, I am able to examine differences within laypeople which previous studies usually could not achieve or did not take into consideration.

Canonicity

Canonicity refers to prototypical or standard forms. Questions can be classified as canonical or non-canonical based on either their form or their function.

Canonical questions regarding function, as e.g. Trotzke (2023) defines them, ask for information. The speaker assumes the addressee possesses the information and is willing to share it (cf. Trotzke, 2023: 23). Non-canonical questions, in contrast, do not seek new information. Trotzke assumes that these functional differences often manifest in linguistic features such as word order or modal particles (cf. Trotzke, 2023: 11).

Concerning form, non-canonical questions do not follow standard rules for question formation in the respective language. In German, canonical questions follow two criteria:

1. They are prosodically (most commonly through rising intonation) or graphemically (question mark) marked.
2. They either have a special word order (V1-order, where the sentence starts with the finite verb), start with a question word (*wh*-word) or end with a question tag (cf. Trotzke, 2023: 11).

Non-canonical questions regarding form have a different lexico-grammatical form (second criterion not fulfilled) but are still marked as a question.[2] In this study, the term (*non-*)*canonical question* refers to form-based canonicity while exploring its relationship to function, hypothesising that non-canonical forms correlate with non-information-seeking functions. Figure 1 gives an overview of canonical questions types in German and non-canonical questions types identified in my corpus data.

Knowledge and Epistemic Status

In Interactional Linguistics, especially in Conversation Analysis, questions and their interpretations in interaction are looked at from the angle of epistemic status. *Epistemic status* refers to the amount of knowledge a person has about a topic at one moment in time (cf. Heritage, 2012: 4). The *epistemic gradient* is the difference

[2] In this study, I ignore speech acts with a questioning function that do not have a final question mark. I consider marking an utterance with a question mark as conscious choice from the commenter that guides my definition of *question*. Starting from the linguistic form of questions, I adopt a form-to-function-approach (cf. Landert et al. 2023: 5), whereas many studies in CA pursue a function-to-form approach (cf. Stivers 2010, Heritage 2012, Bongelli et al. 2018).

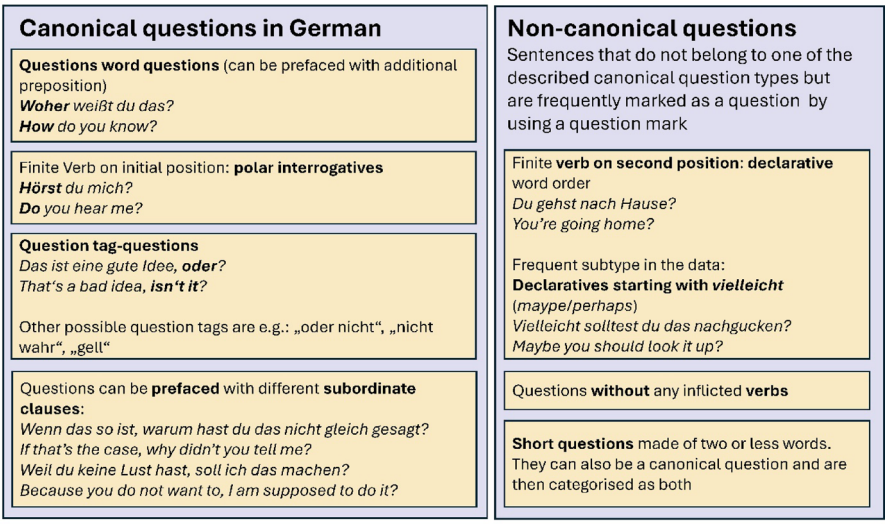

Fig. 1 Overview: canonical question types in German and non-canonical question types identified in this studies' corpus

between the statuses of interlocutors (cf. Heritage, 2012: 6). In contrast, the *epistemic stance* is the expression of the epistemic status:

> In general, of course, unknowing speakers ask questions […], and knowing speakers make assertions. Thus we may speak of a principle of epistemic congruency in which the epistemic stance encoded in a turn at talk will normally converge with the epistemic status of the speaker relative to the topic and the recipient. However, […] this realization is far from inevitable. Epistemic status can be dissembled by persons who deploy epistemic stance to appear more, or less, knowledgeable than they really are. (Heritage, 2012: 7)

While discussants cannot modify the epistemic status they possess, they can modify the epistemic stance they take. One way of taking epistemic stances is through linguistic behaviour, e.g. the form of questions. Linguistic analysis allows access to the epistemic stance but not to the epistemic status.

> A further consequence is that the epistemic stance generally conveyed by interrogative morphosyntax and intonation functions as a secondary lamination on to epistemic status, fine tuning the epistemic gradient between speaker and recipient. (Heritage, 2012: 24)

The form of a question can therefore be a tool for "fine-tuning" the epistemic relation displayed between questioner and addressee.

Bongelli et al. 2018 build on Heritage's model. They differentiate Heritage's *K-* (=little knowledge)-position (cf. Heritage, 2012: 7) into *U* (Unknowing), *NKW* (Not

Knowing Whether), and three degrees of *B* (Believing) (Figure 2). The slope of the arrows represents the epistemic gradient between questioner and answerer.

Examples for the question types taken from this papers' corpus data are:

(1) **Q1** Wer wäre denn jetzt der Ansprechpartner in unserer Bundesregierung? (https://scilogs.spektrum.de/gedankenwerkstatt/die-reiter-der-apokalypse/#comment-26163)
'Who would be the contact person in our federal government now?'

(2) **Q2** a) Ist meine Berechnung richtig? (https://scilogs.spektrum.de/himmelslichter/komet-ison-mit-dem-teleobjektiv-aber-nicht-im-teleskop/#comment-2356)
'Is my calculation correct?'

(3) **Q2** b) Oder wurden Sie etwa[3] von Herrn Dueck beauftragt, hier gleich den Hilfsheriff zu spielen? (https://scilogs.spektrum.de/wild-dueck-blog/stigmenwechsel/#comment-14044)
'Or were you ordered by Mr Dueck to play the deputy sheriff here?'

(4) **Q3** So genau können Satelliten Oberflächenwinde ja gar nicht messen, oder? (https://scilogs.spektrum.de/klimalounge/werden-tropenstuerme-schlimmer/#comment-15393)
'Satellites can't measure surface winds that accurately, can they?'

(5) **Q4** Du testest eine neue Reisemontierung? (https://scilogs.spektrum.de/himmelslichter/morgenstund/#comment-2329)
'You're testing a new travel mount?'

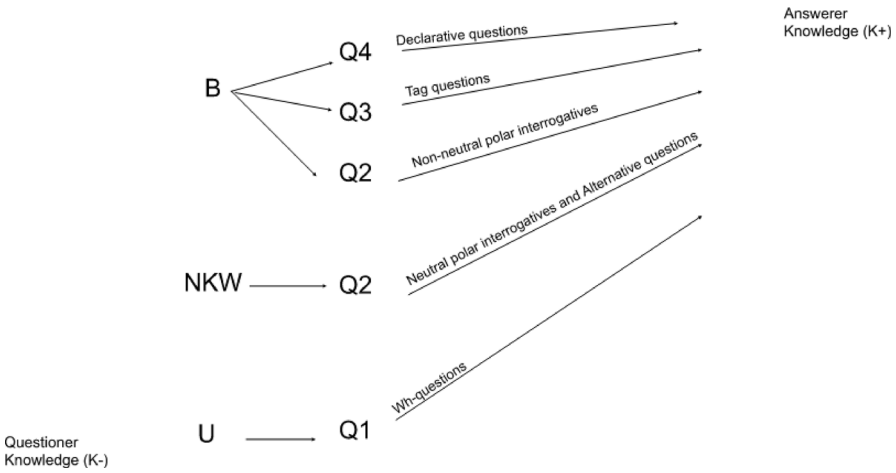

Fig. 2 Visualisation of the epistemic gradient (Bongelli et al., 2018: 42)

[3] The modal particle *etwa* expresses that the questioner hopes that the assertion is not true (cf. Thurmair 1989: 171).

Some questions, like *wh*-questions, imply a bigger difference in knowledge between questioner and addressee.[4] Other questions, like tag questions, show that the questioner sees the discussants as close in regard to epistemic status.

Expertise is associated with knowledge and a presumably higher epistemic status. Enacting expertise involves adopting a higher epistemic stance. The differences among the question types should be detectable in the data if they are linked to specific epistemic constellations.

Politeness and *Face*

This study focuses on questions whose functions extend beyond information seeking. Why do people use questions if they are not lacking information? One of the reasons to present utterances in the form of a questions is politeness and face keeping. I use Brown & Levinson's concept of *face* (Brown & Levinson, 2016 [original work 1987]) as a theoretical background to analyse such question use.

Their *face*-concept assumes the existence of an abstract positive and negative face that each individual has and wants to keep intact. The positive face is the desire to be approved of (e.g. that others perceive what you do or have as good). The negative face is the need for freedom (e.g. freedom of action) (cf. Brown & Levinson, 2016: 61–62).

Actions can threaten either the speaker's or addressee's face; for example, telling someone what to do harms their negative face. Such utterances are called face-threatening acts (FTAs) (cf. Brown & Levinson, 2016: 65). The face damage can be counteracted in several ways. One possibility is "redressive action" (Brown & Levinson, 2016: 69) such as modifications in formulation or additional linguistic embedding (cf. Brown & Levinson, 2016: 68–69). Phrasing statements as questions could minimise the harm, especially to the negative face, since questions leave room for more acceptable reactions than declaratives or imperatives. This strategy matches Brown & Levinsons description of "other softening mechanisms that give the addressee an 'out', a face-saving line of escape, permitting him to feel that his response is not coerced" (2016: 70).

More generally without mention of the *face*-concept, Trotzke also states that turning statements into questions reduces the potential for conflict and disagreement (2023: 37). The hypothesis deriving from this is that formulating an utterance as a question can be a discursive strategy to minimise harm or disagreement by keeping an out-option for the addressee.

To sum up and connect the previous sections: Formulating utterances as questions is a tool that fulfils many different functions. Some of these functions, and therefore potentially also forms, are more likely to be found in experts' comments and others more likely to be found in laypeople's comments.

[4] The studied data consists of public online comments and therefore has multiple addressees: the blogger, potentially the commenter to whose comment the questioner reacts, other commenters, and the imagined audience (cf. Marwick & Boyd 2011: 115–116). In this study, I include all those actors and concepts in the term *addressee*.

Methodology and Methods

Methodological Presuppositions: Corpus Pragmatics, Linguistic Patterns, and Quantitative Methods

Studying the connection of forms and communicative functions falls within the scope of pragmatics: the study on how utterances derive meaning in context and perform communicative action (cf. Rühlemann & Aijmer, 2015: 2). While corpus pragmatic studies integrate quantitative analysis of corpora with pragmatic research, this methodological approach appears to create tension with traditional pragmatic studies, which typically employ qualitative analysis focused on detailed contextual examination. Corpus pragmatic analysis started by combining *horizontal reading*, the qualitative analysis of texts in context, with *vertical reading*, such as reading concordances, and sometimes complementing it with annotation of features (cf. Rühlemann & Aijmer, 2015: 9–12). Quantitative approaches aim to detect recurring linguistic patterns that can be interpreted as a result of pragmatic routines. Specific patterns appearing repeatedly on the linguistic surface likely have pragmatic functions within discourse (cf. Bubenhofer, 2009: 53, Scharloth, 2018: 65). A data-driven approach to corpora, using methods such as keyword-, n-gram-, or collocation analysis, enables the identification of potential patterns. I employ further statistical analyses, such as regression analyses. These analyses allow for the combination and correlation of different phenomena and their relationship to other (extralinguistic) factors, helping to measure their relative importance and to determine their functions and pragmatic meaning potential.

Statistical analysis in corpus linguistics has evolved significantly, moving beyond basic significance testing to more sophisticated methodological approaches. While modern corpus linguistics introductions (Hirschmann, 2019, Stefanowitsch, 2020) address statistical testing, they typically focus on simpler methods rather than complex statistical modelling. Comprehensive statistical introductions for linguists (Gries, 2013; Winter, 2019) primarily demonstrate advanced modelling using experimental data from phonetics or psycholinguistics, leaving a gap in application to pragmatics.

Some recent corpus linguistics papers[5] use advanced statistical methods, but these typically investigate factors influencing linguistic choices (Zhang & Xu, 2023, Zhang & Yue, 2024). In contrast, this research reverses the analytical perspective by examining how linguistic patterns predict non-linguistic outcomes. I supplement the regression analyses with annotation to verify assumptions about reasons for non-canonical question usage (detailed workflow in Section "Workflow").[6]

[5] published in the International Journal of Corpus Linguistics

[6] LLMs or other neural machine learning is not suitable for this analysis. By nature, machine learning can only be used for prediction and cannot inform about causal relationships. In contrast, results from regression analysis can be used to quantify the relative importance of different variables and thereby enable the investigation of underlying mechanisms.

Regression Analysis

The goal of the empirical analysis is to investigate whether different types of questions are used differently by experts or laypeople.

Regression analysis is a statistical method that examines relationships between variables by modelling how multiple independent variables predict a dependent variable. This analytical approach enables researchers to quantify both the combined and relative effects of various predictors on an outcome of interest.

In this study, regression analysis is employed to determine the probability that a given comment was authored by either an expert or a layperson based on specific linguistic features and question characteristics within the comments. The analyses identify variables that are statistically typical for comments from one group compared to comments written by individuals outside that group.

Because the data I want to predict in this study is binary (true or false) I employ logistic regression (cf. Winter, 2019: 198). From a corpus pragmatic perspective, these logistic regressions effectively "create" distinct subcorpora. One subcorpus comprises all comments belonging to the respective group of discourse participants (dependent variable), whilst the remaining comments serve as the reference corpus for comparison. This approach of data-driven comparison between different (sub) corpora as a corpus pragmatic method, as described by Scharloth and Bubenhofer (2012: 199), is enhanced through regression analysis as it allows the test for statistical significance and the control of multiple influencing factors.

Workflow

Figure 3 visualises the workflow and methodological approaches of this paper.

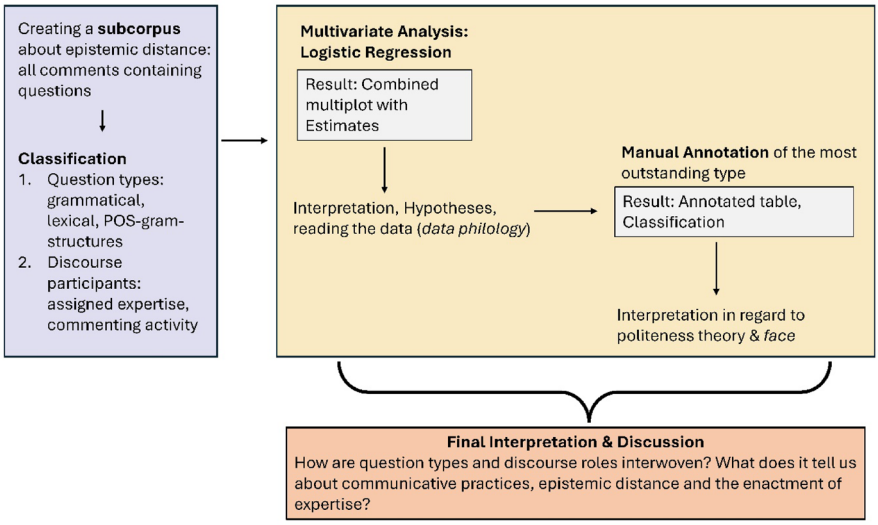

Fig. 3 Steps of analysis and interpretation

I start with creating a subcorpus including all comments containing at least one question. I classify the comments regarding their authorship and the linguistic characteristics of the questions. Logistic regressions identify correlations between groups and comment characteristics, followed by visualisation and interpretation of the regression results. Manual annotation of one particularly salient question type serves to verify hypotheses and provide deeper insights into question functionalities. The combined findings reveal how questions function in epistemic positioning and expertise enactment.

Corpus Data

The general project corpus (Bender et al., 2024) consists of all blog posts and their respective comments on the German science blogging platform *SciLogs—Tagebücher der Wissenschaft*[7] ('diaries of science'), which belongs to the publishing house *Spektrum*.

The project corpus consists of all 267,247 texts published on the platform between 2000 and 2024. These include blogs and their comments, covering diverse research fields, including sciences like astrophysics, chemistry, and biology, but also humanities and social sciences such as theology, psychology, and linguistics. The data is POS-tagged using the Stuttgart-Tübingen-Tag-Set (STTS) and contains rich metadata. Bender et al., 2024 describe the preprocessing and the metadata of the corpus in detail (193–194).

For the purpose of this study, I focus solely on comments and therefore exclude the blog posts. The comments are sentence-tokenised[8]. Emojis pose a problem for this tokenisation since they are frequently used instead of punctuation. Therefore, several sentences are tokenised as one sentence if they lack punctuation due to emojis or non-standard writing.

The final data set includes all comments containing at least one sentence-final question mark. By doing this, I ignore other speech acts with a questioning function. False positives can occur when commenters accidentally use question marks.

I exclude comments that contain mainly words tagged as non-German (POS-tag *FM*). Another challenge are questions that are direct quotes from earlier comments and should therefore not be interpreted as characteristic linguistic feature for the current discourse participant. Therefore, if a question appeared previously in the conversation, it is marked as a duplicate and excluded from the analysis.

These steps lead to a question-focused subcorpus of 77,912 comments, about a third of all 245,796 comments.

This dataset is particularly suitable for researching epistemic status and expertise, as the bloggers demonstrate unusually high engagement in comment sections. While experts comment little on other platforms, *SciLogs* presents an environment where experts contribute substantially to discussions. This allows for direct comparison between expert and lay commenters' comments, rather than merely contrasting

[7] https://scilogs.spektrum.de/

[8] I used the *NLTK sentence tokenizer* (https://www.nltk.org/api/nltk.tokenize.html).

the language in experts' science communication content with laypeople's comments. Additionally, since *Spektrum* verifies the experts who write on the platform, I can rely on the institutional verification rather than developing criteria for expert classification.

The corpus does not aim to be representative of German internet language or language in comment sections. The communicative settings in comment sections vary across platforms. On *Scilogs*, the commenting community is relatively small, and there is reduced anonymity compared to other platforms, as bloggers are identifiable by name. Previous research by Bender et al., using n-gram analyses, has demonstrated notably high levels of enacted politeness in the comments (Bender et al., 2024).

Corpus quotations retain original grammatical and orthographic errors without [sic]-notation.

Classifying the Data—Creating Variables

The quantitative analyses require a classification of discourse participants on the one hand and linguistic properties of the comments on the other hand.

Discourse Roles on *SciLogs*: Experts and Laypeople?

The first set of variables includes the different groups of discourse participants on the platform.[9] I use this classification to access the epistemic implications of question types by analysing the correlation of types with different discourse participant groups.

Commenters are divided into four groups, based on two criteria: the quantity of the commenting activities and whether they are the author of the respective blog post (cf. Fig. 4).

The bloggers are the experts for the topics evoked from the blog post. My presupposition for the analysis is that the bloggers are the institutionalised experts, whereas the other commenters are not. This rigid assignment might seem in opposition to Carr's theory of enactment of expertise. Of course, some of the other commenters can also be experts, but the platform itself institutionally frames the bloggers as experts. The bloggers successfully enacted their expertise to *SciLogs* and are now perceived as experts. Their blogs give them voice and agency. In terms of conversation analysis, it is also important to bear in mind that they are automatically the person who has the first turn by writing the blog article. All comments are subsequent reactions to the initial article. If bloggers write a comment on a different blog than their own, they are not categorised as bloggers since expert roles are relative.

The second classification criterion is the quantity of comments written. One group consists of individuals who commented only once or twice. This group functions as a "control group" without a distinctive commenting style since they probably are a heterogeneous group that only comments occasionally with a concrete question

[9] Sex and gender (SAGER) variables are not relevant for this study and therefore SAGER guidelines are not applicable.

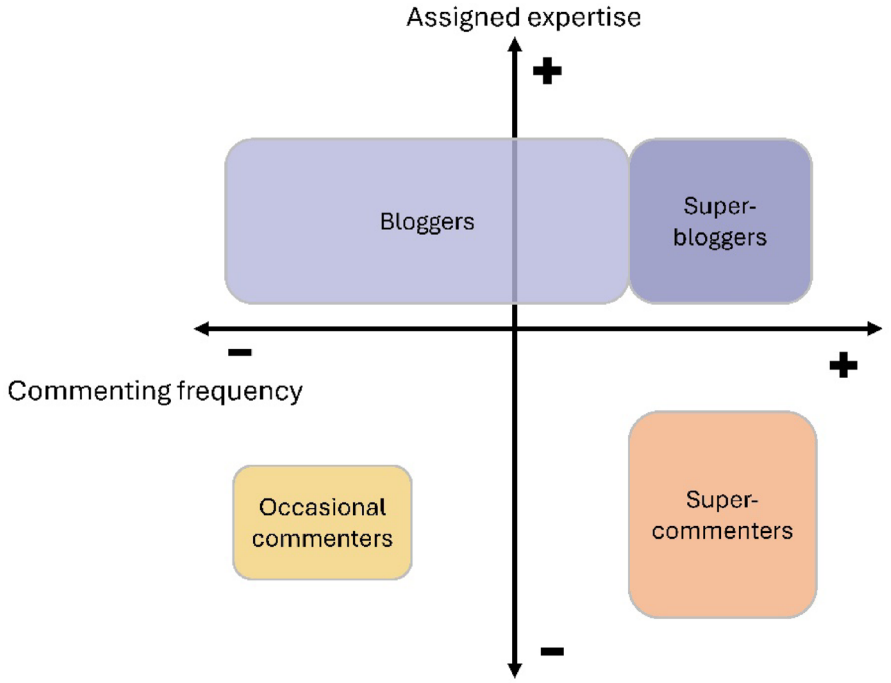

Fig. 4 Groups of categorised discourse participants

Table 1 Number of members and written comments of the different groups of discourse participants

Group	Group members (in subcorpus)	Comments (whole corpus)	Comments (subcorpus)
Bloggers without super-bloggers	167 (65)	7,255	1,411
Occasional commenters	14,367 (4,595)	16,528	4,910
Super-commenters without super-bloggers	126 (123)	120,563	40,169
Super-bloggers	25 (25)	35,806	10,430

or remark in mind, but not because they want to discuss or negotiate expertise and science. This group includes 14,367 people who contributed 16,528 comments (cf. Table 1).

Another group, super-commenters, represents a small but highly active segment of the commenting community. To qualify as a super-commenter, an individual must have commented on more than 200 articles across at least 10 different SciLogs-blogs, demonstrating broad engagement with scientific discussions. While comprising only 0.7% of all commenters (126 out of 18,132), they contribute nearly half of all comments across the platform. They are anonymous users who create their online persona, gain discursive power, and enact their expertise. Some super-commenters are so present in the comments that they evoke meta-discursive utterances where they are described as *seltsame Akteure* ('strange actors') and *übliche Verdächtige* ('usual suspects'); their commenting practices are portrayed as *ständig ausufernde Kaperung der Kommentare* ('constantly excessive hijacking of comments') that drives other

commenters away from *SciLogs* (cf. discussion underneath this article https://scilogs.spektrum.de/wild-dueck-blog/stigmenwechsel).

The last group consists of "super-bloggers", who are both bloggers and super-commenters. Their dual role will show whether the way of writing questions is more influenced by expert status or by commenting activity. They are excluded from the blogger and super-commenter categories to prevent skewed results and ensure accurate interpretation.

123 super-commenters wrote about 40,000 of the around 78,000 comments in the subcorpus of this study. The 25 super-bloggers wrote 10,500 comments.

There are some commenters who are not part of any of the groups (people who wrote between 3 and 197 comments and are not bloggers). These individuals are included in the control group for all analyses as I conduct separate regression analyses for each group so that each group is compared to non-members of that specific group, rather than directly with other groups.

Because authors usually belong to one group, it is not possible to statistically separate author effects from group effects since the inclusion of authors as random effect variable would override all possible linguistic features.

While the groups vary in size, this does not compromise the validity of the analysis. Larger groups typically yield narrower confidence intervals, while smaller groups are more likely to show statistically nonsignificant results.

Question Types and Other Linguistic Patterns

The second set of variables is made by classifying some of the linguistic properties of the comments. For the question types, I built a question classifier for German questions.[10] Due to the above mentioned issues with automated tokenisation some questions are not or wrongly classified because the question is not found within the wrongly tokenised "sentence". Precision and Recall values for all categories can be found in the Appendix.

The classification began with identifying canonical questions like polar interrogatives (V1-order), question-word questions, tag questions, and declarative questions (V2-order without question tags).

Additional patterns were identified through data-driven analysis of remaining unclassified questions with frequency lists and n-gram analysis. This inductive approach is suitable for discovering patterns and non-canonical question types used in online comments.

The classifier assigns Boolean values to each question type at the comment level. Categories are non-exclusive. A single comment may contain multiple questions, and a single questions may fall into multiple categories (e.g. both *wh*-word and short questions).

The classification (Table 2) derives directly from corpus data. As online comments tend towards conceptually oral language, the questions often deviate from standard forms.

[10] Data, code, and visualisations of the results as html-plots are available as online resources.

Table 2 Question types, examples and their frequencies. The sum of the relative frequency does not add up to 100% as a question can be categorised to be part of more than one type and comments can contain multiple different questions.

Type	Example	Absolute number (relative amount) of comments containing type
Question-word question (*Wh*-word question)	*Wieso fragen Sie das?* 'Why do you ask that?'	33,125 (42.52%)
Polar interrogative	*Kann man das so simpel hochrechnen?* 'Is it possible to scale it up that easily?'	28,826 (37.00%)
Question tag question	*Könnte doch auch sein,* oder nicht *?* 'Could be, couldn't it?'	2,953 (3.80%)
Vielleicht('maybe'/'perhaps')-question	*Vielleicht sollten Sie eine diesbezügliche Anfrage an die zuständige PTB Arbeitsgruppe 4.41 richten?* 'Perhaps you should ask the PTB working group 4.41?'	1,531 (1.97%)
Question that starts with a *wenn*('if')-clause	*Wenn die Welt sowieso untergeht, warum dann nicht noch wenigstens etwas Spaß haben?* 'If the world is coming to an end anyway, why not at least have some fun?'	3,700 (4.75%)
Short question of max. 2 words	*Na und?* 'So what?'	9,238 (11.86%)
Declarative question (V2-order without other question marker)	*Sie sagen selbst, dass RKI «Mistzahlen» liefert?* 'You say yourself that RKI gives "rubbish numbers"?'	7,317 (9.39%)
Question without inflicted verb	*Heimat großer Töchter und Söhne?* 'Home to great daughters and sons?'	10,087 (12.95%)
Comments with only unclassified questions	/	11,338 (14.55%)

The classification includes three additional linguistic features common in questions: the modal verb *sollen* ('should'), multiple questions marks and personal addressing of the question.

Sollen (particularly in *Konjunktiv II*) indicates that the content of the utterance represents someone else's assumption or expectation (cf. https://www.dwds.de/wb/sollen#d-1-3-2). It creates distance from or devalues the statement's content, suggesting scepticism. In terms of epistemic gradient, this implies the questioner claims superior knowledge by questioning the validity of the statement. Using multiple question marks indicates heightened expressivity and emotional involvement. Personal addressing suggests a predetermined recipient or participation in multi-party discussions where specific addressees need identification (examples and numbers in the Appendix).

An additional variable is the position of the question within the comment and the number of questions (Table in Appendix). Comment-initial questions could, e.g. tend to have a rhetorical function and let the commenter answer their question in subsequent text. Final questions could be the result of prior reflection in the comment. The

number of questions in the comment as absolute and relative values can be revealing since it tells something about the shape of the comment.

By combining all these variables, it is possible to estimate the differences between the groups of discourse participants and show differences in linguistic behaviour concerning questioning practices in online comments.

Combination of the Variables: Regression Analysis

The results of the analyses are visualised as coefficient plots (also available as html-plots in this repository: https://doi.org/10.48656/vs99-an46). All graphs are created with *plotly R* [11], the coefficient tables of all regressions are given in the Appendix. A coefficient plot shows statistical relationships through dots (point estimates) and horizontal bands (confidence intervals 95%). When a dot appears above zero, the feature makes it more likely that a comment comes from the respective group. When it appears below zero, the feature decreases that probability. The horizontal band indicates the certainty of the estimate: if this band crosses the zero-line, the relationship is not statistically significant. If it stays entirely on one side of the zero-line, the relationship can be considered non-random.

One problem concerning regressions with many predictors is that with the number of predictors the probability of committing a type I error (i.e. falsely rejecting a true null hypothesis) increases. To test the robustness of the results, several Tables the Appendix document the regression results with Bonferroni-Holm-Correction where the p-values are adjusted so that they take the multiple-comparison problem into account. All relevant results remain significant.

Some question types show significant differences between groups, whereas other types are insignificant for all groups of discourse participants.

The first three variables in Figure 5 are canonical questions. *Wh-word questions* and polar interrogatives do not seem to be especially (un-)typical for any of the groups. If a comment contains a tag question, the probability that a super-commenter wrote it is increased, whereas it is atypical for the super-bloggers. Tag questions have two general functions: They can be understood "(a) as a way of requesting information, normally confirmation of the assertion made in the declarative component of the utterance, or (b) as a method for mobilizing response" (Heritage, 2012: 14). They imply a small epistemic difference since they refer to non-certainty about a subject rather than to not-knowing (cf. Bongelli et al., 2018: 31, Simon & Janich, 2021: 26).

The five variables thereafter correspond to non-canonical question types. The highest inter-group variability is visible for the *vielleicht* ('maybe'/'perhaps')-questions. They are the most outstanding result for both groups of bloggers. Figure 6 shows the predicted values of the probability of the comment being written by bloggers or super-bloggers depending on this variable.

If the comment does not contain a *vielleicht*-question, the probability is around 4%/12%, whereas if the comment does contain one, the probability rises to around

[11] https://plotly.com/r/

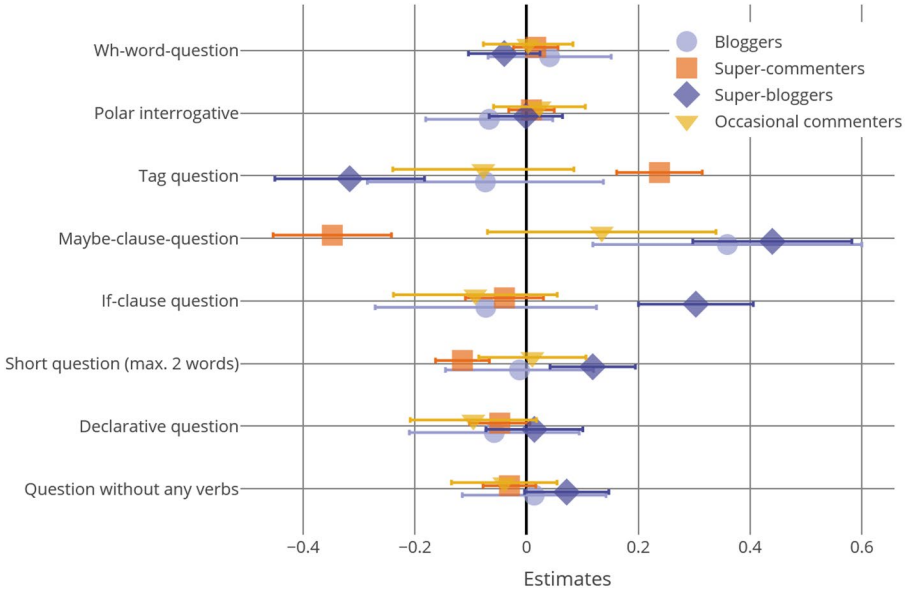

Fig. 5 Multiplot with results of regression analyses (log odds) for all four groups of discourse participants for all question types

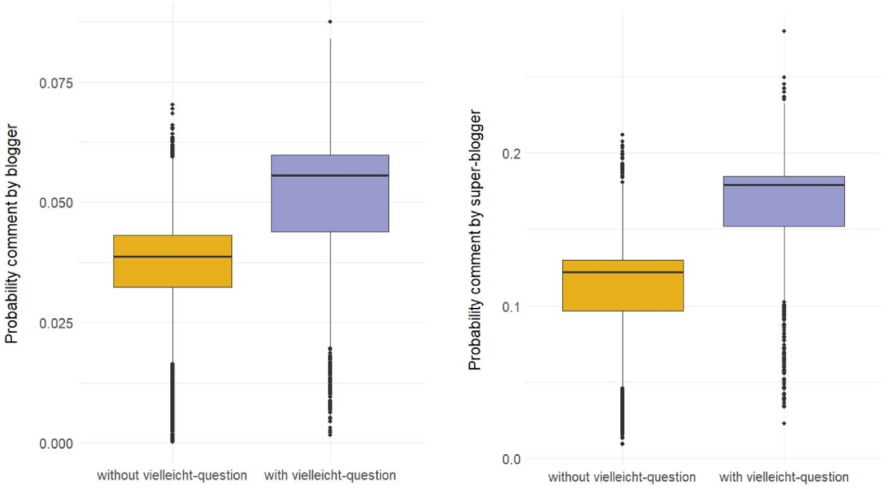

Fig. 6 Predicted values for the variable *Vielleicht*-questions for the bloggers and super-bloggers

6%/18%. This shows how the use of *vielleicht*-questions augments the probability by approximately a third for both blogger groups compared to without this question.

Wenn-questions show a tendency to be used by the super-bloggers. This pattern appears connected to the function of *wenn* ('if'/'when') to react to previous statements. Super-bloggers show a higher tendency to engage with existing comments rather than initiating new discussions, which makes sense given that the blog posts

already serve as initial turn in the discussion. The frequent use of *wenn*-clauses prefacing questions suggests a particular communicative strategy: challenging or questioning the validity of previous statements.

(6) Wenn zutrifft, was Sie meinen – warum haben dann die Macher der Studie Wert darauf gelegt, dass die Sticker-Empfänger "den gleichen ethnischen Hintergrund" wie die Geber haben sollten? (https://scilogs.spektrum.de/natur-des-glaubens/sind-kinder-eine-studie-fehler/#comment-82583)
'If what you say is true, why did the creators of the study emphasise that the sticker recipients should have 'the same ethnic background' as the givers?'

(7) @[username]: Wenn es so selten passiert, halten Sie dann die angebliche Milliarde, die Sitzenbleiber jährlich kosten, für plausibel? (https://scilogs.spektrum.de/menschen-bilder/sitzenbleiben-ist-sinnlos/#comment-30894)
'@[username]: If it happens so rarely, do you think the alleged billions that dropouts cost every year are plausible?'

Short questions are correlated with super-bloggers' comments and are not typical of super-commenters' comments. Declaratives that generally hint at a smaller epistemic distance are rather atypical for occasional commenters. Questions that do not contain any tagged verbs are not typical for either group.

Beyond question types, several additional linguistic patterns significantly influence the likelihood of comment authorship across different groups (Figure 7).

Comments that start with a question tend to be written by the groups who do not comment frequently, whereas the "super-groups" tend to not start their comments with a question. Prefacing questions with additional sentences gives the questioner the possibility to explain relevance and adapt the context for the question (cf. Clayman & Heritage, 2002:193, 195–196).

No group shows a tendency to close their comment with a final question; occasional commenters explicitly tend not to do this.

Personal addressing exhibits a negative correlation with low-frequency users and a strong positive correlation with super-commenters and super-bloggers. This suggests that most of their comments come from discussions with one another, where questions are directed not to the reading public but at the other discussant.

The modal verb *sollen* 'should' is slightly typical for both expert groups. Rare commenters tend not to use it. This specific usage in questions is linked to expertise and to the discourse power to criticise and challenge assumptions (cf. "Question Types and Other Linguistic Patterns").

(8) Wie soll es denn dafür empirische Befunde geben, wie sollten die denn aussehen? (https://scilogs.spektrum.de/menschen-bilder/von-terroranschlaegen-zur-willensfreiheit/#comment-55751)
'How are there supposed to be empirical findings for this, what should they look like?'

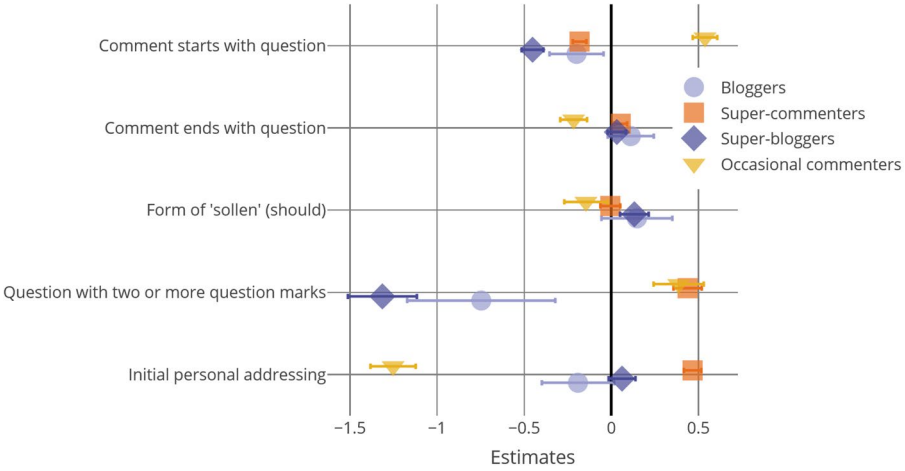

Fig. 7 Multiplot with results of regression analyses (log odds) for all four groups of discourse participants for the position of the question inside the comment and other linguistic features of questions

(9) Sie scheinen unbewusst anzunehmen, dass die Evolution "rationales" Verhalten besonders begünstigen sollte. Doch warum sollte das so sein? (https://scilogs.spektrum.de/natur-des-glaubens/wenn-gott-leid-menschen-von/#comment-58537)
'You seem to unconsciously assume that evolution should particularly favour 'rational' behaviour. But why should this be the case?'

Most of the time, *sollen* appears in combination with a *wh*-word.

Using multiple question marks is atypical for the experts, whereas the "non-expert"-commenters use this graphemic feature more frequently. It puts expressive emphasis on the questions marked with it and suggests emotional involvement and therefore also unprofessional communicative behaviour (examples 10 and 11). Using several question marks could be a strategy to create voice in discussions in the absence of formal assigned expertise. Frequently, several questions also emphasise a question's rhetorical nature.

(10) Was geht denn hier ab? Ist das noch SciLogs?? (https://scilogs.spektrum.de/natur-des-glaubens/gruppenbezogene-menschenfeindlichkeit-gmf-in-evolution-rer-perspektive/#comment-24457)
'What's going on here? Is this still SciLogs??'

(11) Woher kommt bloß Ihre Auffassung, dass eine öffentliche kontroverse Debatte dazu dienen sollte, die Protagonisten jeweils von ihren gegenteiligen Meinungen zu überzeugen??? (https://scilogs.spektrum.de/datentyp/ki-aufruf-an-die-politik-gastbeitrag/#comment-3294)
'Where did you get the idea that a public controversial debate should serve to convince the protagonists of their opposing opinions???'

The analysis concludes by examining how the relative and absolute frequency of questions within comments correlate with the groups of authors (Figure 8).

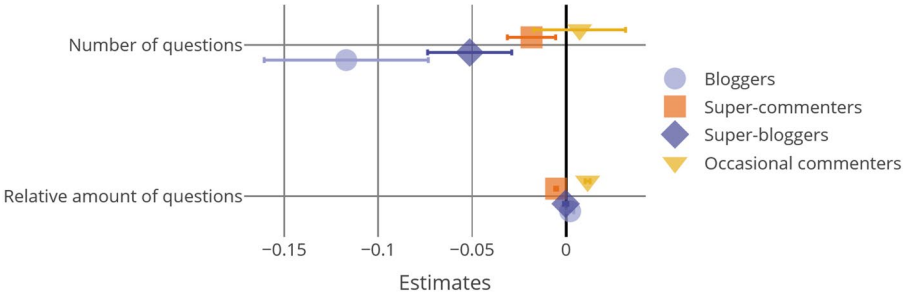

Fig. 8 Multiplot with results of regression analyses (log odds) for all four groups of discourse participants for number and amount of questions

All groups, except the occasional commenters, tend to use few questions. The probability that bloggers or super-commenters wrote a comment decreases with each additional question included. This pattern suggests a straightforward interpretation: fewer questions may indicate less need for information, potentially reflecting greater expertise (or perceived expertise). While questions can serve various epistemic functions beyond information-seeking, the presence of many canonical, information-seeking questions in the data supports this interpretation.

The relative proportion of questions in a comment strongly correlates with overall comment length, measured by the total number of sentences. The analysis reveals that as the percentage of questions (relative to non-question sentences) increases, the probability of the comment originating from an occasional commenter rises. Conversely, this probability decreases for super-commenters. This pattern suggests two distinct commenting behaviours: occasional commenters typically write shorter comments centred around specific queries, indicating that their primary purpose is to seek information rather than engage in extended discussions. In contrast, super-commenters tend to craft lengthier contributions with a proportionally lower question density, suggesting more elaborate argumentative or explanatory discourse patterns.

To conclude, the statistical analysis reveals distinct questioning patterns characteristic of different participant groups:

Bloggers demonstrate a preference for non-canonical questions. The significant prevalence of *vielleicht*-questions warrants further investigation through functional annotation (cf. Section "Closer Look at Possible Functions: Annotating Vielleicht-Questions"). The frequent use of the modal verb *sollen* emphasises the critical-interrogative nature of their contributions.

Super-bloggers' questioning patterns align more closely with bloggers than super-commenters, suggesting that expert status exerts a stronger influence than community engagement levels. However, super-bloggers distinctively demonstrate more frequent direct addressing of comments and use *wenn*-questions, a feature not significant for the other bloggers.

Super-commenters, as highly active participants, exhibit different linguistic behaviours. They frequently employ tag questions, which reduce epistemic distance, whilst using fewer canonical and *vielleicht*-questions. Their comments often are extensive, with questions comprising a relatively small proportion of the total content.

A marked distinction exists between super-commenters and occasional commenters, revealing significant variation within the lay participant category. Occasional commenters typically initiate their contributions with questions and maintain greater epistemic distance in their questioning style. They write concise comments, often limited to one or several questions.

It is important to emphasise that these findings indicate tendencies rather than exclusive patterns of usage. The results demonstrate statistical preferences rather than absolute group boundaries.

Closer Look at Possible Functions: Annotating *Vielleicht*-Questions

To verify the functionality of the statistically most outstanding question type, I conduct functional annotation of the *Vielleicht*-questions.

The annotation was a two-step process. In the first round of annotation, I developed the functional categories in a hermeneutic reading process of 100 random comments, including *vielleicht*-questions, initially resulting in ten categories. Following reliability testing with two annotators, these were consolidated into five core categories to reduce overlap and improve inter-annotator reliability. In a second round, 100 new random comments were annotated using these final categories:

1. Answer suggestion: Propositions for potential answers or solutions, directed at specific interlocutors, the general audience, or as self-reflection.
2. Directive Functions: Requests or suggestions for action, excluding propositions for verbal responses.
3. Hypothesis Formation: Expressing tentative assumptions or speculations about possible explanations or outcomes.
4. Request for information, confirmation, or clarification.
5. Rest: Sentences that are wrongly assigned as questions due to an sentence-internal question mark (example 12) or other problems during sentence tokenization (cf. Section "Classifying the Data – Creating Variables"), or consisted of only one *Vielleicht* lacking sufficient context for functional categorisation.

(12)Vielleicht kann Teilchenbeschleunigung (?) das ja lösen.
'Perhaps particle acceleration (?) can solve this problem'

In 68 of 100 cases the annotators agreed but the agreement differed according to the different categories (Table 3):

The highest inter-annotator reliability was achieved for the directive function. The information-seeking questions were rarely interpreted in consent, which might be due to the general rarity of this category. This is another hint to the tendency that questions in non-canonical forms tend to fulfil non-canonical functions.

The high value of Cohen's Kappa and the generally frequent occurrence of *vielleicht*-questions with a directive function makes them particularly suitable for in-depth analysis.

Using *vielleicht,* the questioner directs the addressee (one person, a group, or unspecified) to take an action. Making demands inherently threatens the addressee's negative face. The inclusion of *vielleicht* functions as a linguistic mitigation device, whilst the interrogative form provides addressees with greater perceived freedom of response. This dual strategy of using both *vielleicht* and the question format creates a double mitigation of the FTA.

However, when looking at some occurrences, it becomes apparent that it can also be used in a destructive, sometimes even sarcastic way, like in examples 13 and 14:

(13)Vielleicht wäre es ein Start, überhaupt mal zu lesen, was das von dir genannte Paper überhaupt aussagt? (https://scilogs.spektrum.de/klimalounge/2000-jahre-meeresspiegel/#comment-6845)
'Maybe it would be a good start to read what the paper you mentioned actually says?'

(14)Vielleicht gelingt Ihnen ja doch noch ein Ausweg aus Ihrem Hass und Verschwörungsglauben? (https://scilogs.spektrum.de/natur-des-glaubens/verschwoerungsfragen-9-adolf-hitler-als-rothschild-verschwoerer-der-libertaere-antisemitismus/#comment-104039)
'Perhaps you can still find a way out of your hatred and belief in conspiracies?'

In these occurrences, the *Vielleicht…?* pattern cannot be interpreted as mitigation. When applied to deliberately insulting demands, the meaning potential of *Vielleicht…?* can be twisted to serve ironic or hostile functions, indicating its status as a recognisable pragmatic pattern in the discourse community.

Discussion and Outlook: What do the Results Show About the Construction of Expertise?

In comments on *SciLogs*, expertise and epistemic stance can be constructed through questions. By studying how different groups of discourse participants act linguistically, different questioning patterns become apparent. These usage patterns can indicate the pragmatic and epistemic value of question types. Questions typically employed by experts demonstrate distinctive characteristics: they likely exhibit reduced epistemic distance and deviate from conventional information-seeking questions. The regression analyses demonstrate that non-canonical question forms tend to be used by commenters who are less likely to use questions to gain new information. The non-canonical forms appear to be linked to non-canonical functions.

From these findings, I develop an expansion to Heritage's and Bongelli's models of epistemic hierarchies in questioning to account for two more possible epistemic hierarchies between questioner and addressee while reducing the former levels to three (Figure 9).

Table 3 Annotation Categories with examples and the Inter Annotator Agreement (Cohens Kappa)

Function	Example with some context	Agreed/ only A1/ only A2	Co-hen's Kappa
Answer suggestion	*Würdest du das tun? Und wenn nicht , warum dann nicht? Vielleicht, weil dieser Klon eben doch nich du selbst wäre?* 'Would you do that? And if not, why not? Maybe because this clone wouldn't be you after all?'	13/21/2	0.407
Directive Functions	*Vielleicht könnten Sie eine statistische Analyse oder ein Paper zur Unterstützung der Behauptung mitliefern?* 'Perhaps you could provide a statistical analysis or a paper to support the claim?'	26/2/5	0.832
Hypothesis Formation	*Vielleicht geht die wirtschaftliche Entwicklung einmal wieder zu diesem Stadium zurück , wenn der letzte Monopolist Konkurs gegangen ist?* 'Perhaps economic develop-ment will return to this stage once the last monopolist has gone bankrupt?'	24/3/15	0.599
Request for information	*Viele Grüße und bis bald, spätestens bis zum nächsten Bloggertreffen; vielleicht sieht man sich aber vorher auf einer Konferenz?* 'Many greetings and see you soon, at the latest by the next blog get-together; but maybe we'll see each other at a conference before then?'	1/5/7	0.08
Rest	See example 12	4/1/3	0.646

Question types cannot be pinned down to one arrow, but in tendency, the researched question types are expected to be somewhat close to the following hierarchy-constellations:

1. Canonical *wh*-question
2. V1-questions (mostly polar interrogatives)
3. Declarative questions and tag questions
4. Some of the *vielleicht*-questions (those suggesting answers), *wenn*-questions, and also declarative and tag questions if they possess additional features like certain modal particles or several question marks
5. Directive *vielleicht*-questions and *wh*-questions containing *sollen*, that flip the conventional questioning hierarchy

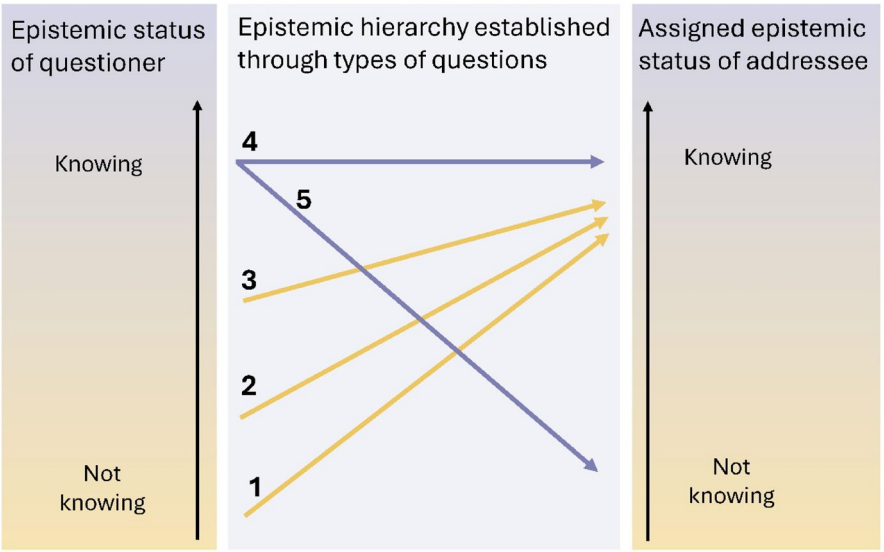

Fig. 9 Model of possible epistemic constellations

The model includes the notion that questions do not invariably signal a lower epistemic status of the questioner. This becomes particularly evident in challenging or doubting questions, where questioning itself suggests higher knowledge and critical engagement with the discussed subject. This observation helps explain why experts, despite their presumed greater knowledge, employ questions in discourse.

The key to understanding this phenomenon lies in how questions operate on multiple levels. At the surface level, questions inherently protect the addressee's face by appearing to grant them higher epistemic status. This built-in face protection creates a pragmatic buffer that enables speakers to express critical or doubtful perspectives while maintaining a polite tone. By using some questions like the *vielleicht*- or *wenn*-questions, the potentially critical content can be mitigated through the question form.

The bloggers, as discourse participants with the highest assigned expertise, frequently employ these question types. This group's comments are likely to contain these questions and features to claim a higher epistemic position than the addressee.

While the bloggers and the super-bloggers share many characteristics, the laypeople show substantial differences in their questioning behaviour. The occasional commenters fulfil the 'stereotypical' role of laypeople by asking canonical questions with presumably canonical functions and thereby keeping the epistemic hierarchy intact. The use of several question marks shows emotional involvement and potentially the need to create voice in this discursive sphere that few people dominate.

The super-commenters also use questions that minimise epistemic distance, like tag questions. However, unlike bloggers, they use a different form of question (tag question instead of *vielleicht*-question) to reach a similar communicative goal (e.g. suggesting an answer more indirectly and politely). Their linguistic behaviour appears to be shaped not only by them being (very interested) laypeople, but also by their belonging to this subgroup of commenters who pursue commenting as a

hobby, who partly also like to provoke controversial discussions, and seek intellectual stimulation.

Different linguistic features characterise the groups of discussants. The supercommenters do not aim to copy the experts' questioning patterns to simulate their way of enacting expertise. They use their own linguistic patterns to discuss and construct expertise.

This paper's data and approach offer the unique opportunity to observe the commenting practices of experts. Typically, in other online discussions, the experts are not recognisable through institutional authorisation. These clearly assigned roles on *SciLogs* enable a direct comparison between experts and laypeople. On other platforms every discussant has to enact their expertise to gain discursive power. Investigating which of the above described questioning strategies are employed on other platforms would be an important next step towards understanding how people construct expertise in the internet more generally. I assume that users on other media platforms also use questions to display expertise, and I have shown several possibilities how this can be achieved.

Enactment of expertise happens when questioners put themselves in higher epistemic positions than the addressees. While canonical questions typically signal knowledge gaps or uncertainty, certain questioning strategies establish the questioner's authority.

To capture these interactions the model could include the corresponding answers. Fully understanding questioning practices necessitates examination of answering patterns. Which questions receive responses, and what patterns are detectable in them? How do different groups approach answering? For some examples in their qualitative conversation analysis, Bongelli et al., 2018 already include the answers and add them to their model (Figure 10). How to assess the epistemic stances in answers quantitatively needs further investigation.

On a methodological level, I demonstrated that regression analysis is an effective tool for identifying significant and non-significant linguistic patterns across participant groups. This analytical approach revealed distinctive questioning practices for each group, and in combination, it gives an insight into the commenting practices of the different discourse participants. The analytical process followed three steps:

1. Data-driven classification of features on the linguistic surface,
2. Application of regression analyses across several groups,
3. Comparison of the results of the regressions to identify group-specific characteristics.

Regression analysis offers particular value in linguistic research through its capacity to evaluate multiple variables simultaneously while assessing their relative impact on outcomes. This approach enables researchers to compare the influence of linguistic features against extra-linguistic factors, potentially revealing that linguistic variables carry equal or greater weight than other assumed influential factors.

The successful application of this analytical approach requires careful consideration of both dependent and independent variables. A particular challenge lies in identifying and defining appropriate dependent variables that are both measurable

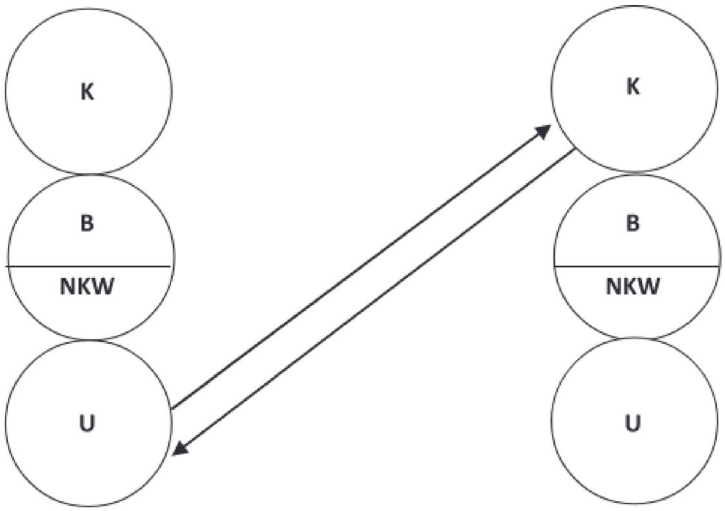

Fig. 10 Question-answer sequence (Bongelli et al., 2018: 32)

and meaningful. While this necessity for predefined variables might appear to conflict with *data-driven* approaches, regression analyses can be effectively integrated at various stages of the research process. Initial exploratory 'purely data-driven' analyses like key word- or n-gram-analyses can inform variable selection, which leads to a hypothesis that is then studied further with statistical methods.

Conclusion

Questions have the potential to be used as a tool to enact expertise. Certain question types and linguistic patterns inside questions challenge the conventional epistemic relationship between questioner and addressee. Both experts and super-commenters employ distinctive questioning strategies to position themselves epistemically, necessitating an expansion of traditional epistemic hierarchy models to include new possible configurations.

The bloggers enact their expertise through distinctive use of *vielleicht*-questions and *sollen*, the super-bloggers additionally use *wenn*-questions. Super-commenters, in contrast, use tag questions, which establish less pronounced epistemic hierarchies than *wh*-questions. They also show a significant tendency towards personal addressing. Notably, the super-bloggers align more often with regular bloggers than

super-commenters, suggesting that expert status influences linguistic behaviour more strongly than frequency of participation .

Occasional commenters display different patterns, showing barely any significant affinity for non-canonical question types. Significant characteristics are the use of several question marks and not addressing the questions. These differences to super-commenters' comments highlight that different laypeople act differently and cannot be seen as one homogenous group united by their non-expert status.

Questions can express expertise and superior knowledge while minimising face threats. Non-canonical question forms, in particular, serve complex pragmatic functions beyond information-seeking. I show the tendency for questions in non-canonical forms to fulfil non-canonical functions.

Regression analysis is an innovative methodological tool in corpus pragmatics. The approach proved effective in identifying statistically significant patterns of question usage across different discourse participant groups, revealing subtle and pronounced linguistic variations that might not be immediately apparent through traditional analytical methods. The combination of different analyses of several groups enables the detection of less salient linguistic patterns that carry significant pragmatic weight in questioning behaviour. While regression analysis offers new possibilities for corpus pragmatic research, its application requires careful consideration of the variables, as identifying and quantifying appropriate variables may be difficult. Despite this challenge, the methodology demonstrates considerable promise for future research in corpus pragmatics.

Given the exploratory approach of this study, it is important to verify the generalisability of the result that expertise is enacted through certain types of questions. The methods should be applied to other data where groups with different epistemic statuses interact.

This study opens several promising avenues for further investigation into questioning practices in scientific discourse: Further linguistic analysis such as n-gram analysis could reveal more nuanced patterns within question types. Such fine-grained investigations could identify more specific lexical or syntactic features that correlate with different *vielleicht*-questions, which would enable the identification of more precise form-meaning relationships than the current broader categorisation permits. To better understand the functions of non-canonical questions, two helpful approaches could be:

1. Systematic annotation of additional non-canonical question types to better understand their interactive functions, and
2. Analysis of metalinguistic discourse about questioning practices within the community, particularly focusing on how participants discuss and evaluate "(non-) genuine" questions.

This study demonstrates how questions serve as tools for enacting expertise while simultaneously revealing the complex interplay between linguistic form and pragmatic function in scientific discourse.

Appendix

Classification Tables

Tables 4 and 5

Table 4 Linguistic features included in the classifier

Type	Example	Absolute number (relative amount) of comments that contain feature
Form of modal verb *sollen* ('should')	*Warum **sollte** denn Religiösität nur über einen Nachkommen weitergegeben werden?* 'Why should religiosity only be passed on through a descendant?'	5,513 (7.08%)
Multiple question marks	*Weshalb also nicht jetzt schon "Staatssprache Englisch"??* 'So why not 'English as the national language' now??'	2,640 (3.39%)
Personal addressing	@[username], *warum versuchen sie mir einen Strohmann unterzuschieben?* '@[username], why are you trying to foist a straw man [argument] on me?'	8,862 (11.37%)

Table 5 Properties of the comments

Feature	Statistics
Comment starts with a question	17,357 (22.28%)
Comment ends with a question	16,632 (21.35%)
Absolute number of questions in the comment	Median: 1, Mean: 1.88
Relative amount of sentences in a comment that are questions (in %)	Median: 16.67, Mean: 24.15

Error Analysis Question Classifier

Most categories' classifications have high Precision and Recall values and only few false classifications. The classification of declarative questions poses most problems as the category has many false positives due to problems in sentence tokenisation Table 6.

Table 6 Results of manual error analysis of the question classifier. True Positives, True Negatives, False Positives and False Negatives for all categories and the resulting Metrics (Precision, Recall, Accuracy, F-measure)

Category	TP	TN	FP	FN	Precision	Recall	Accuracy	F_measure
wh question	70	67	2	10	0.972	0.875	0.919	0.921
preposition + wh question	11	136	1	1	0.917	0.917	0.987	0.917
polar interrogative	56	81	4	8	0.933	0.875	0.919	0.903
tag question	11	137	1	0	0.917	1.0	0.993	0.957
short question	30	116	2	1	0.938	0.968	0.980	0.952
if-clause question	18	130	0	1	1.0	0.947	0.993	0.973
because-clause question	10	138	1	0	0.909	1.0	0.993	0.952
maybe-clause question	10	139	0	0	1.0	1.0	1.0	1.0
form of sollen	20	127	1	1	0.952	0.952	0.987	0.952
declarative question	12	124	8	5	0.600	0.706	0.913	0.649
no verbs	24	120	5	0	0.828	1.0	0.966	0.906

Coefficient Tables with Bonferroni-Holm-Correction, AIC and Pseudo R^2

Tables 7, 8, 9 and 10

Table 7 Regression results bloggers: Estimates, p-values and p-values with Bonferroni-Holm-correction. Due to the Bonferroni-Holm-correction in comparison to the main specification two variables become insignificant for the group of the bloggers: question starts with a question and the number of questions

Variable	Estimate	Std. Error	p-val	Bonf.-Holm p-val	Signif. p-val	Signif. Bonf.-Holm p-val
(Intercept)	-3.829935	0.073183	0.000000	0.000000	True	True
is initialTRUE	-0.199444	0.079106	0.011695	0.163728	True	False
is finalTRUE	0.112043	0.067280	0.095845	1.000000	False	False
wh questionTRUE	0.009650	0.078182	0.901770	1.000000	False	False
polar interrogativeTRUE	-0.030832	0.080416	0.701418	1.000000	False	False
tag questionTRUE	-0.168250	0.157750	0.286171	1.000000	False	False
maybe clause questionTRUE	0.506823	0.160470	0.001587	0.023798	True	True
if clause questionTRUE	-0.096086	0.142176	0.499153	1.000000	False	False
short questionTRUE	0.107986	0.091810	0.239519	1.000000	False	False
declarative questionTRUE	-0.100279	0.109519	0.359864	1.000000	False	False
no verbsTRUE	0.004756	0.092492	0.958986	1.000000	False	False
unclassified questionTRUE	-0.081104	0.106024	0.444298	1.000000	False	False
form of sollenTRUE	0.147031	0.103772	0.156525	1.000000	False	False
sev questionmarksTRUE	-0.746426	0.216805	0.000576	0.009210	True	True
initial addressingTRUE	-0.191314	0.105653	0.070176	0.842108	False	False
question count	-0.061483	0.028571	0.031401	0.408207	True	False
question percentage	0.000008	0.001428	0.995727	1.000000	False	False

AIC: 14083

Table 8 Regression results super-commenters: Estimates, p-values and p-values with Bonferroni-Holm-correction. The variable unclassified questions becomes insignificant in comparison to the main specification due to the correction. This variable is not interpreted

Variable	Estimate	Std. Error	p-val	Bonf.-Holm p-val	Signif. p-val	Signif. Bonf.-Holm p-val
(Intercept)	0.183929	0.019431	0.000000	0.000000	True	True
is initialTRUE	-0.182210	0.019958	0.000000	0.000000	True	True
is finalTRUE	0.054016	0.018771	0.004007	0.032054	True	True
wh questionTRUE	0.032497	0.020343	0.110167	0.661004	False	False
polar interrogativeTRUE	0.003648	0.020800	0.860778	1.000000	False	False
tag questionTRUE	0.275474	0.039351	0.000000	0.000000	True	True
maybe clause questionTRUE	-0.315149	0.053562	0.000000	0.000000	True	True
if clause questionTRUE	-0.011829	0.035627	0.739873	1.000000	False	False
short questionTRUE	-0.117749	0.024544	0.000002	0.000016	True	True
declarative questionTRUE	-0.024124	0.027559	0.381363	1.000000	False	False
no verbsTRUE	-0.036586	0.024041	0.128057	0.661004	False	False
unclassified questionTRUE	0.056161	0.027680	0.042467	0.297269	True	False
form of sollenTRUE	-0.005582	0.029041	0.847588	1.000000	False	False
sev questionmarksTRUE	0.438074	0.041259	0.000000	0.000000	True	True
initial addressingTRUE	0.466652	0.025570	0.000000	0.000000	True	True
question count	-0.020258	0.006546	0.001970	0.017734	True	True
question percentage	-0.005274	0.000382	0.000000	0.000000	True	True

AIC: 107116

Table 9 Regression results super-bloggers: Estimates, p-values and p-values with Bonferroni-Holm-correction. The variable short question becomes insignificant due to the correction

Variable	Estimate	Std. Error	p-val	Bonf.-Holm p-val	Signif. p-val	Signif. Bonf.-Holm p-val
(Intercept)	-1.684172	0.028640	0.000000	0.000000	True	True
is initialTRUE	-0.451840	0.031753	0.000000	0.000000	True	True
is finalTRUE	0.031915	0.026825	0.234154	1.000000	False	False
wh questionTRUE	-0.026075	0.030736	0.396251	1.000000	False	False
polar interrogativeTRUE	-0.020404	0.031539	0.517670	1.000000	False	False
tag questionTRUE	-0.271171	0.062879	0.000016	0.000178	True	True
maybe clause questionTRUE	0.416573	0.069188	0.000000	0.000000	True	True
if clause questionTRUE	0.261420	0.049990	0.000000	0.000002	True	True
short questionTRUE	0.085025	0.036789	0.020825	0.187423	True	False
declarative questionTRUE	0.007278	0.041599	0.861122	1.000000	False	False
no verbsTRUE	0.066699	0.036117	0.064784	0.518273	False	False
unclassified questionTRUE	0.023485	0.040900	0.565822	1.000000	False	False
form of sollenTRUE	0.132490	0.041883	0.001560	0.015599	True	True
sev questionmarksTRUE	-1.313744	0.100979	0.000000	0.000000	True	True
initial addressingTRUE	0.061733	0.039284	0.116083	0.812584	False	False
question count	-0.067435	0.011037	0.000000	0.000000	True	True
question percentage	0.000657	0.000549	0.231769	1.000000	False	False

AIC: 60693

Table 10 Regression results occasional commenters: Estimates, p-values and p-values with Bonferroni-Holm-correction. The variable form of *sollen* becomes unsignificant due to the correction of the p-values

Variable	Estimate	Std. Error	p-val	Bonf.-Holm p-val	Signif. p-val	Signif. Bonf.-Holm p-val
(Intercept)	-2.997027	0.040154	0.000000	0.000000	True	True
is initialTRUE	0.537828	0.035857	0.000000	0.000000	True	True
is finalTRUE	-0.216011	0.039191	0.000000	0.000000	True	True
wh questionTRUE	0.002769	0.040889	0.946005	1.000000	False	False
polar interrogativeTRUE	0.022983	0.041898	0.583313	1.000000	False	False
tag questionTRUE	-0.077461	0.082708	0.348982	1.000000	False	False
maybe clause questionTRUE	0.134481	0.104341	0.197449	1.000000	False	False
if clause questionTRUE	-0.091787	0.074805	0.219814	1.000000	False	False
short questionTRUE	0.010288	0.048974	0.833614	1.000000	False	False
declarative questionTRUE	-0.095327	0.057668	0.098327	0.983275	False	False
no verbsTRUE	-0.040011	0.048199	0.406470	1.000000	False	False
unclassified questionTRUE	-0.020551	0.056461	0.715860	1.000000	False	False
form of sollenTRUE	-0.145405	0.063373	0.021766	0.239422	True	False
sev questionmarksTRUE	0.386709	0.073465	0.000000	0.000002	True	True
initial addressingTRUE	-1.252314	0.066270	0.000000	0.000000	True	True
question count	0.007272	0.012486	0.560294	1.000000	False	False
question percentage	0.011525	0.000670	0.000000	0.000000	True	True

AIC: 35785

Correlation and Collinearity: Correlation Plot of All Variables Used in the Regressions and Variance Inflation Factor (VIF)

To investigate whether the variables are correlated, I use a correlation plot to visualise the correlations between all variables (Fig. 11). Additionally, I report the Variance Inflation Factor which measures collinearity between variables (Tables 11, 12, 13, and 14).

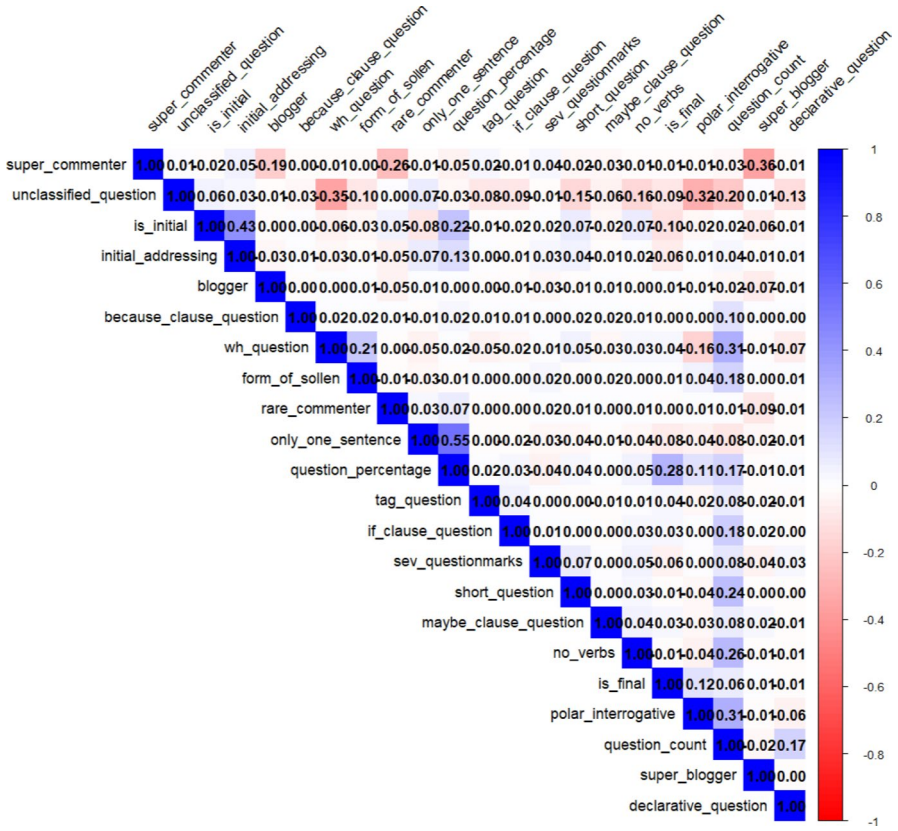

Fig. 11 Correlation plot of all variables used in the regressions. The number of questions (question_count) slightly correlates with some values which is not surprising since comments that contain many questions are more likely to include many different question types

Table 11 Variance Inflation Factor bloggers

Variable	VIF-value
is initial	1.274737
is final	1.135208
wh question	2.072320
polar interrogative	2.054245
tag question	1.066692
maybe clause question	1.087801
if clause question	1.105570
short question	1.229584
declarative question	1.255469
no verbs	1.262203
unclassified question	1.847472
form of sollen	1.069326
sev questionmarks	1.009058
initial addressing	1.198768
question count	2.102667
question percentage	1.198048

Table 12 Variance Inflation Factor super-commenters

Variable	VIF-value
is initial	1.318283
is final	1.139314
wh question	1.947204
polar interrogative	1.941373
tag question	1.064663
maybe clause question	1.046329
if clause question	1.105072
short question	1.207795
declarative question	1.244618
no verbs	1.251218
unclassified question	1.831099
form of sollen	1.067670
sev questionmarks	1.017801
initial addressing	1.240445
question count	2.125091
question percentage	1.204287

Table 13 Variance Inflation Factor super-bloggers

Variable	VIF-value
is initial	1.283287
is final	1.121754
wh question	2.055460
polar interrogative	2.047511
tag question	1.062066
maybe clause question	1.070193
if clause question	1.135001
short question	1.222033
declarative question	1.274514
no verbs	1.268316
unclassified question	1.893915
form of sollen	1.069462
sev questionmarks	1.006092
initial addressing	1.230675
question count	2.154301
question percentage	1.176465

Table 14 Variance Inflation Factor: occasional commenters

Variable	VIF-value
is initial	1.190398
is final	1.143868
wh question	1.853378
polar interrogative	1.878714
tag question	1.055432
maybe clause question	1.046757
if clause question	1.092755
short question	1.200042
declarative question	1.208030
no verbs	1.243030
unclassified question	1.803327
form of sollen	1.059263
sev questionmarks	1.022694
initial addressing	1.090322
question count	2.018529
question percentage	1.214864

Regressions with Firth Correction

Tables 15, 16, 17 and 18

Table 15 Regression results bloggers with Firth correction

Variable	Estimate	Std.Error	pval	signif
(Intercept)	-3.829498	0.072886	0.000000	True
is initialTRUE	-0.198444	0.078819	0.010905	True
is finalTRUE	0.112382	0.067070	0.096973	False
wh questionTRUE	0.008764	0.077846	0.910770	False
polar interrogativeTRUE	-0.031658	0.080063	0.693745	False
tag questionTRUE	-0.158262	0.156500	0.303135	False
maybe clause questionTRUE	0.516249	0.159244	0.002416	True
if clause questionTRUE	-0.088959	0.141194	0.525565	False
short questionTRUE	0.109853	0.091406	0.235310	False
declarative questionTRUE	-0.097406	0.108991	0.369078	False
no verbsTRUE	0.006458	0.092063	0.944312	False
unclassified questionTRUE	-0.080347	0.105659	0.447078	False
form of sollenTRUE	0.150551	0.103299	0.153229	False
sev questionmarksTRUE	-0.724837	0.213806	0.000144	True
initial addressingTRUE	-0.187969	0.105152	0.069877	False
question count	-0.060138	0.028364	0.029450	True
question percentage	0.000047	0.001423	0.973927	False

Table 16 Regression results super-commenters with Firth correction

Variable	Estimate	Std.Error	pval	signif
(Intercept)	0.183892	0.019428	0.000000	True
is initialTRUE	-0.182163	0.019955	0.000000	True
is finalTRUE	0.053997	0.018769	0.004006	True
wh questionTRUE	0.032447	0.020338	0.110550	False
polar interrogativeTRUE	0.003601	0.020794	0.862538	False
tag questionTRUE	0.275282	0.039340	0.000000	True
maybe clause questionTRUE	-0.314933	0.053539	0.000000	True
if clause questionTRUE	-0.011865	0.035617	0.739047	False
short questionTRUE	-0.117752	0.024538	0.000002	True
declarative questionTRUE	-0.024160	0.027553	0.380602	False
no verbsTRUE	-0.036614	0.024036	0.127740	False
unclassified questionTRUE	0.056121	0.027676	0.042550	True
form of sollenTRUE	-0.005601	0.029035	0.847042	False
sev questionmarksTRUE	0.437776	0.041246	0.000000	True
initial addressingTRUE	0.466498	0.025566	0.000000	True
question count	-0.020219	0.006541	0.001909	True
question percentage	-0.005273	0.000382	0.000000	True

Table 17 Regression results super-bloggers with Firth correction

Variable	Estimate	Std.Error	pval	signif
(Intercept)	-1.684155	0.028626	0.000000	True
is initialTRUE	-0.451605	0.031737	0.000000	True
is finalTRUE	0.031964	0.026815	0.234093	False
wh questionTRUE	-0.026199	0.030719	0.393903	False
polar interrogativeTRUE	-0.020530	0.031521	0.515004	False
tag questionTRUE	-0.269841	0.062813	0.000010	True
maybe clause questionTRUE	0.417669	0.069117	0.000000	True
if clause questionTRUE	0.261883	0.049954	0.000000	True
short questionTRUE	0.085221	0.036768	0.021191	True
declarative questionTRUE	0.007540	0.041574	0.856189	False
no verbsTRUE	0.066845	0.036096	0.065148	False
unclassified questionTRUE	0.023523	0.040882	0.565332	False
form of sollenTRUE	0.132884	0.041859	0.001706	True
sev questionmarksTRUE	-1.309071	0.100738	0.000000	True
initial addressingTRUE	0.062030	0.039262	0.115664	False
question count	-0.067236	0.011026	0.000000	True
question percentage	0.000661	0.000549	0.230274	False

Table 18 Regression results occasional commenters with Firth correction

Variable	Estimate	Std.Error	pval	signif
(Intercept)	-2.996599	0.040111	0.000000	True
is initialTRUE	0.537727	0.035828	0.000000	True
is finalTRUE	-0.215797	0.039156	0.000000	True
wh questionTRUE	0.002247	0.040822	0.956164	False
polar interrogativeTRUE	0.022436	0.041827	0.592191	False
tag questionTRUE	-0.075211	0.082531	0.358193	False
maybe clause questionTRUE	0.138169	0.104041	0.191930	False
if clause questionTRUE	-0.090219	0.074658	0.222730	False
short questionTRUE	0.010525	0.048907	0.829895	False
declarative questionTRUE	-0.094824	0.057580	0.097416	False
no verbsTRUE	-0.039862	0.048129	0.406840	False
unclassified questionTRUE	-0.020621	0.056400	0.714705	False
form of sollenTRUE	-0.144126	0.063278	0.020688	True
sev questionmarksTRUE	0.388381	0.073335	0.000000	True
initial addressingTRUE	-1.250465	0.066166	0.000000	True
question count	0.007747	0.012429	0.537986	False
question percentage	0.011528	0.000669	0.000000	True

Regressions Outcomes with Different Classifications of Super-Commenters and Super-Bloggers

To verify whether the categories are robust, I calculated the regressions of the super-

Table 19 Regression results super-commenters with a threshold of 100: Estimates, p-values and p-values with Bonferroni-Holm-Correction

term	estimate	std error	pval	pval bonf	signif	signif bonf
(Intercept)	0.312521	0.019471	0.000000	0.000000	True	True
is initialTRUE	-0.143054	0.019932	0.000000	0.000000	True	True
is finalTRUE	0.051280	0.018818	0.006428	0.051420	True	False
wh questionTRUE	0.018799	0.020375	0.356169	1.000000	False	False
polar interrogativeTRUE	0.001894	0.020833	0.927567	1.000000	False	False
tag questionTRUE	0.282812	0.039775	0.000000	0.000000	True	True
maybe clause questionTRUE	-0.252880	0.053098	0.000002	0.000019	True	True
if clause questionTRUE	-0.032352	0.035668	0.364391	1.000000	False	False
short questionTRUE	-0.120013	0.024529	0.000001	0.000011	True	True
declarative questionTRUE	-0.026677	0.027600	0.333768	1.000000	False	False
no verbsTRUE	-0.032465	0.024062	0.177266	1.000000	False	False
unclassified questionTRUE	0.034054	0.027765	0.220010	1.000000	False	False
form of sollenTRUE	-0.025355	0.029101	0.383596	1.000000	False	False
sev questionmarksTRUE	0.419445	0.041778	0.000000	0.000000	True	True
initial addressingTRUE	0.404352	0.025701	0.000000	0.000000	True	True
question count	-0.024240	0.006538	0.000209	0.001884	True	True
question percentage	-0.004387	0.000380	0.000000	0.000000	True	True

commenters and super-bloggers with different thresholds of 100 comments and 300 comments instead of 200 comments. I keep the criterion that they have to have com-

Table 20 Regression results super-commenters with a threshold of 300: Estimates, p-values and p-values with Bonferroni-Holm-Correction

term	estimate	std error	pval	pval bonf	signif	signif bonf
(Intercept)	0.083877	0.019454	0.000016	0.000162	True	True
is initialTRUE	-0.221997	0.020063	0.000000	0.000000	True	True
is finalTRUE	0.048671	0.018796	0.009615	0.086534	True	False
wh questionTRUE	0.052266	0.020371	0.010295	0.086534	True	False
polar interrogativeTRUE	0.019789	0.020825	0.341984	1.000000	False	False
tag questionTRUE	0.280978	0.039158	0.000000	0.000000	True	True
maybe clause questionTRUE	-0.330238	0.054096	0.000000	0.000000	True	True
if clause questionTRUE	0.005921	0.035669	0.868157	1.000000	False	False
short questionTRUE	-0.117914	0.024620	0.000002	0.000018	True	True
declarative questionTRUE	-0.028652	0.027601	0.299221	1.000000	False	False
no verbsTRUE	-0.036765	0.024092	0.127011	0.635053	False	False
unclassified questionTRUE	0.066609	0.027684	0.016124	0.097896	True	False
form of sollenTRUE	-0.001293	0.029047	0.964490	1.000000	False	False
sev questionmarksTRUE	0.422834	0.040810	0.000000	0.000000	True	True
initial addressingTRUE	0.473883	0.025490	0.000000	0.000000	True	True
question count	-0.016112	0.006556	0.013985	0.097896	True	False
question percentage	-0.006179	0.000385	0.000000	0.000000	True	True

Table 21 Regression results super-bloggers with a threshold of 100: Estimates, p-values and p-values with Bonferroni-Holm-Correction

term	estimate	std error	pval	pval bonf	signif	signif bonf
(Intercept)	-1.663655	0.028423	0.000000	0.000000	True	True
is initialTRUE	-0.444327	0.031471	0.000000	0.000000	True	True
is finalTRUE	0.043625	0.026599	0.100985	0.706898	False	False
wh questionTRUE	-0.035227	0.030513	0.248305	1.000000	False	False
polar interrogativeTRUE	-0.023179	0.031319	0.459257	1.000000	False	False
tag questionTRUE	-0.252884	0.061985	0.000045	0.000496	True	True
maybe clause questionTRUE	0.447339	0.068173	0.000000	0.000000	True	True
if clause questionTRUE	0.257745	0.049702	0.000000	0.000003	True	True
short questionTRUE	0.102912	0.036397	0.004692	0.042227	True	True
declarative questionTRUE	0.000587	0.041372	0.988680	1.000000	False	False
no verbsTRUE	0.069149	0.035846	0.053720	0.429760	False	False
unclassified questionTRUE	0.016119	0.040629	0.691565	1.000000	False	False
form of sollenTRUE	0.129266	0.041696	0.001934	0.019339	True	True
sev questionmarksTRUE	-1.302495	0.099635	0.000000	0.000000	True	True
initial addressingTRUE	0.059265	0.038981	0.128417	0.770505	False	False
question count	-0.068756	0.010980	0.000000	0.000000	True	True
question percentage	0.000664	0.000545	0.223697	1.000000	False	False

Table 22 Regression results super-commenters with a threshold of 300: Estimates, p-values and p-values with Bonferroni-Holm-Correction

term	estimate	std error	pval	pval bonf	signif	signif bonf
(Intercept)	-1.697036	0.028768	0.000000	0.000000	True	True
is initialTRUE	-0.467405	0.031990	0.000000	0.000000	True	True
is finalTRUE	0.031972	0.026923	0.235030	1.000000	False	False
wh questionTRUE	-0.026099	0.030860	0.397707	1.000000	False	False
polar interrogativeTRUE	-0.018268	0.031660	0.563936	1.000000	False	False
tag questionTRUE	-0.275498	0.063225	0.000013	0.000145	True	True
maybe clause questionTRUE	0.395878	0.069813	0.000000	0.000000	True	True
if clause questionTRUE	0.260790	0.050168	0.000000	0.000002	True	True
short questionTRUE	0.084029	0.036944	0.022937	0.206435	True	False
declarative questionTRUE	0.000869	0.041816	0.983421	1.000000	False	False
no verbsTRUE	0.063250	0.036292	0.081364	0.569549	False	False
unclassified questionTRUE	0.022734	0.041082	0.580004	1.000000	False	False
form of sollenTRUE	0.130246	0.042062	0.001958	0.019579	True	True
sev questionmarksTRUE	-1.301718	0.100988	0.000000	0.000000	True	True
initial addressingTRUE	0.073656	0.039430	0.061762	0.494095	False	False
question count	-0.065945	0.011058	0.000000	0.000000	True	True
question percentage	0.000699	0.000551	0.204489	1.000000	False	False

mented on at least 10 different blogs. The results are very similar in all groups. Tables 19, 20, 21 and 22.

Acknowledgements I am grateful for support, guidance and inspiring discussions provided by Noah Bubenhofer and Marie-Luis Merten. I want to thank Seraina Betschart for her assistance in coding the question classifier. I would like to thank the anonymous reviewers for their helpful remarks and suggestions. I want to thank the Schweizer Nationalfonds (SNF) and the Deutsche Forschungsgemeinschaft (DFG) for the funding of the project this paper originates from.

Author Contributions G.K. conducted all analyses, wrote all text, and prepared all figures.

Funding Open access funding provided by University of Zurich. The research leading to these results received funding from Schweizerischer Nationalfonds (SNF) under Grant Agreement No. 214084 and Deutsche Forschungsgemeinschaft (DFG) under Grant Agreement No. 517920702.

Data Availability Data: The corpus was created and developed with linguistic annotation and metadata by the DiscourseLab of the TU Darmstadt. The corpus data is stored on the CQPweb instance of the Discourse Lab belonging to the TU Darmstadt (https://www.discourselab.de/cqpweb/scilogs2024/). However, the corpus is not publicly accessible because Spektrum prohibits the distribution of the data due to legal constraints. Individual access rights to the corpus data can be requested via e-mail to: hiwi_dislab@linglit.tu-darmstadt.de. All texts can be separately accessed via: https://scilogs.spektrum.de/ Cited examples can be found via the provided links that lead to the blog articles and the respective discussions. The results of the classification and raw data for the regression is provided in this SWISSUbase repository: https://doi.org/10.48656/9f3v-g730. Code: The question classifier and the R-Script used for statistical analysis is provided in these SWISSUbase repositories: https://doi.org/10.48656/ngkb-dg18, https://doi.org/10.48656/2fdr-3828. Visualisations: The plotly-visualisations as html-files are provided in this SWISSUbase repository: https://doi.org/10.48656/vs99-an46.

Declarations

Conflict of interests The authors declare no competing interests.

Ethical Approval The research only involves publicly available textual data that was not evoked through experiments and therefore does not need ethics approval. Additionally, all quoted corpus examples are anonymised.

References

Antos, G., Niehr, T., & Spitzmüller, J. (Eds.). (2019) *Handbuch Sprache Im Urteil Der Öffentlichkeit*. De Gruyter. https://doi.org/10.1515/9783110296150

Bender, M., Bubenhofer, N. & Janich, N. (2024). 'Die öffentliche Aushandlung von Expertise: Wissenschaftsblogs als Ort eristischer Verständigung? Exploratorischer Einstieg in ein Forschungsprojekt', *Zeitschrift für germanistische Linguistik*, *52*(1), 183–211. https://doi.org/10.1515/zgl-2024-2008

Bender, M., Kessler, D. & Wachter, D. (2024) 'Korpus der "SciLogs – Tagebücher der Wissenschaft" – Wissenschaftsblogs und -kommentierung 2000–2024, CQPWeb-Edition', DiscourseLab Darmstadt

Bock, B., & Antos, G. (2019). ‚Öffentlichkeit' – ‚Laien' – ‚Experten': Strukturwandelvon ‚Laien' und ‚Experten' in Diskursen über ‚Sprache'. In G. Antos, T. Niehr, & J. Spitzmüller (Eds.), *Handbuch Sprache im Urteil der Öffentlichkeit* (pp. 54–80). DeGruyter. https://doi.org/10.1515/9783110296150-004

Bongelli, R., Riccioni, I., Vincze, L., & Zuczkowski, A. (2018). Questions and Epistemic Stance: Some Examples from Italian Conversations. *Ampersand, 5*, 29–44. https://doi.org/10.1016/j.amper.2018.11.001

Breeze, R. (2023). "Not One of Our Experts." Knowledge Claims and Group Affiliations in Online Discussions of the COVID-19 Vaccine. In R. Plo-Alastrué & I.Corona (Eds.), *Digital Scientific Communication: Identity and Visibility in Research Dissemination* (pp. 33–52). Springer International Publishing.https://doi.org/10.1007/978-3-031-38207-9_2

Brown, P., & Levinson, S. C. (2016). *Politeness: Some Universals in Language Usage*, (25th printing 2016). Cambridge University Press.

Bubenhofer, N. (2009). *Patterns of Language Usage. Corpus Linguistics as a Method of Analyzing Discourse and Culture: Korpuslinguistik als Methode der Diskurs- und Kulturanalyse* (Bd. 4). Walter de Gruyter, https://doi.org/10.1515/9783110215854

Carr, E. S. (2010). Enactments of Expertise. *Annual Review of Anthropology, 39*(1), 17–32. https://doi.org/10.1146/annurev.anthro.012809.104948

Clayman, S., & Heritage, J. (2002). *The news interview: Journalists and public figures on the air*. Cambridge University Press.

Coen, S., Meredith, J., Woods, R., & Fernandez, A. (2021). Talk like an Expert: The Construction of Expertise in News Comments Concerning Climate Change'. *Public Understanding of Science, 30*(4), 400–16. https://doi.org/10.1177/0963662520981729

Gries, S. T. (2013) *Statistics for Linguistics with R: A Practical Introduction*, (2. rev. ed). De Gruyter. https://doi.org/10.1515/9783110307474

Heritage, J. (2012). Epistemics in action: Action formation and territories of knowledge. *Research on Language & Social Interaction, 45*(1), 1–29. https://doi.org/10.1080/08351813.2012.646684

Hirschmann, H. (2019) *Korpuslinguistik: Eine Einführung*. J.B. Metzler. https://doi.org/10.1007/978-3-476-05493-7

Hoffmeister, T., Hundt, M., & Naths, S. (Eds.). (2021). *Laien, Wissen, Sprache: Theoretische, methodische und domänenspezifische Perspektiven*. De Gruyter. https://doi.org/10.1515/9783110731958

Landert, D., Dayter, D., Messerli, T. C. & Locher, M. A. (2023). *Corpus Pragmatics*. (Cambridge University Press). https://doi.org/10.1017/9781009091107

Marwick, A. E., & Boyd, D. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society, 13*(1), 114–133. https://doi.org/10.1177/1461444810365313

Merten, M. L. (2024). *Soziale Positionen – soziale Konstruktionen: Stancetaking im Online-Kommentieren*. De Gruyter. https://doi.org/10.1515/9783111530949

Rühlemann, C., & Karin, A. (Eds.). (2015). *Corpus Pragmatics: A Handbook*. Cambridge University Press. https://doi.org/10.1017/CBO9781139057493

Scharloth, J. (2018). Korpuslinguistik für sozial- und kulturanalytische Fragestellungen. In M. Kupietz & T. Schmidt (Eds.), *Korpuslinguistik* (pp. 61–80). De Gruyter. https://doi.org/10.1515/9783110538649-004

Scharloth, J., & Bubenhofer, N. (2012). Datengeleitete Korpuspragmatik. Korpusvergleich als Methode der Stilanalyse. In E. Felder, M. Müller, & F. Vogel (Eds.), *Korpuspragmatik: Thematische Korpora als Basis diskurslinguistischer Analysen* (pp. 195–230). De Gruyter. https://doi.org/10.1515/9783110269574.195

Schrader-Kniffki, M. (2014). Subject Emergence, Self-Presentation, and Epistemic Struggle in French Language Forums. In K. Bedijs, G. Held, & C. Maaß (Eds.), *Face Work and Social Media* (pp. 376–401). Lit-Verl.

Simon, N., & Janich, N. (2021). Fragen und Antworten. Wissenskonstitution in Kontroversen am Beispiel des Glyphosat-Diskurses. *Fachsprache, 43* (1–2), 22–51. https://doi.org/10.24989/fs.v43i1-2.1938

Spitzmüller, J. (2019). ‚Sprache' – , Metasprache' ‚–Metapragmatik': Sprache und sprachliches Handeln als Gegenstand sozialer Reflexion. In G. Antos, T. Niehr, & J. Spitzmüller (Eds.), *Handbuch Sprache Im Urteil Der Öffentlichkeit* (pp. 11–30). De Gruyter. https://doi.org/10.1515/9783110296150

Spitzmüller, J. (2021). His Master's Voice: Die soziale Konstruktion des ‚Laien' durch den ‚Experten'. In T. Hoffmeister, M. Hundt, & S. Naths (Eds.), *Laien, Wissen, Sprache* (pp. 1–24). De Gruyter. https://doi.org/10.1515/9783110731958-001

Stefanowitsch, A. (2020) *Corpus Linguistics: A Guide to the Methodology*. Zenodo. https://doi.org/10.5281/ZENODO.3735822

Stivers, T. (2010). An overview of the question-response system in American English conversation. *Journal of Pragmatics, 42*(10), 2772–2781. https://doi.org/10.1016/j.pragma.2010.04.011

Thurmair, M. (1989). *Modalpartikeln und ihre Kombinationen*. Niemeyer.

Trotzke, A. (2023). *Non-Canonical Questions*. Oxford University Press. https://doi.org/10.1093/oso/9780192872023 2289.001.0001

Winter, B. (2019). *Statistics for Linguists: An Introduction Using R*. Routledge. https://doi.org/10.4324/9781315165547

Zhang, D., & Xu, J. (2023). Dative Alternation in Chinese: A Mixed-Effects Logistic Regression Analysis. *International Journal of Corpus Linguistics, 28*(4), 559–585. https://doi.org/10.1075/ijcl.21086.zha

Zhang, Yi., & Yue, M. (2024). Case and agreement variation in contact: A multifactorial investigation of *It* -clefts across World Englishes. *International Journal of Corpus Linguistics, 29*(4), 472–506. https://doi.org/10.1075/ijcl.22119.zha