Corpus Approaches to Analysing Uncertainty and Ignorance in Academic Discourse

Marcus Müller

Abstract The article provides an overview of corpus approaches to researching linguistic practices for dealing with ignorance and uncertainty. Uncertainty and ignorance are first and foremost epistemological or socio-psychological categories rather than linguistic ones. But they can be applied to a corpus linguistic setting. Based on a presentation of the central terms and their relevance for digital corpus research, this paper exemplifies proposals for operationalisations using a corpus of political science texts from the field of International Relations (DIReC). It gives an overview of various methods of researching ignorance and uncertainty in academic discourse, focusing on lexicon-based, annotation-based and pattern-search-based approaches as well as combinations thereof. The structure of the explanations reflects a central conflict of aims: On the one hand, corpus-based research on ignorance and uncertainty requires a precise, interpretive approach to the contextual meaning and epistemic function of each individual piece of evidence. On the other hand, it seems advantageous to investigate the largest possible corpora for reasons of reliability. The final section presents an application sketch that addresses and exemplifies several methodological problems. It compares uncertainty markers in political science discourse as found in the DIReC corpus with those in conspiracy theories, drawing on the LOCO corpus and journalistic discourse represented in a reference corpus of US newspapers.

Keywords annotation by query, computer linguistics, corpus linguistics, ignorance, machine learning, operationalisation, uncertainty

1 Introduction

The articulation of uncertainty and ignorance is one of the central communicative tasks in academic discourse.[1] While this holds for all academic disciplines, linguistic practices vary across subjects in their extent, distribution and expression. Uncertainty and ignorance can be primarily understood as epistemological or socio-psychological phenomena, but they are reflected in language. Conceptualising these phenomena from a linguistic perspective poses challenges for all empirical linguistic approaches (Janich 2018, 2020, Ratcliff/Wicke/Harvill 2022), but especially for those based on the linguistic-statistical analysis of language data. Nevertheless, digital corpora and algorithmic methods of language analysis can yield new insights into the forms of linguistic expression of uncertainty and ignorance in academic discourse as well as their significance (Jean et al. 2016, Müller/Stegmeier 2019, Zinn/Müller 2021).

[1] I will use *ignorance* and *non-knowledge* synonymously, choosing *ignorance* as a term of my own metalanguage. For a discussion of terminological designation alternatives cf. Japp (2000: 225), Böschen et al. (2010: 784), Nielsen/Sørensen (2017: 386).

Zitiervorschlag / Citation:
Müller, Marcus (2023): "Corpus Approaches to Analysing Uncertainty and Ignorance in Academic Discourse." *Fachsprache. Journal of Professional and Scientific Communication* 45.1–2: 28–47.

This article presents the current state and challenges of research on uncertainty and ignorance in academic discourse in the field of digital linguistics. The linguistic analysis of uncertainty and ignorance relies heavily on contextual information and is therefore a subject of linguistic pragmatics. This makes it especially difficult to identify ignorance-related phenomena in medium and large data sets on the basis of pattern analysis.

I begin this article with an overview of selected core concepts and conceptual preferences – insofar as they are either grounded in digital linguistics or help to discuss digital approaches. I draw on the basic distinction between known unknowns and unknown unknowns and discuss its implications for linguistic operationalisation. Considerations from the fields of philosophy, science studies and the sociology of knowledge that go beyond this limited scope will be therefore left aside.

The subsequent sections present the main methodological problems and challenges in digital research on uncertainty and ignorance and provide an overview of related works. Section 3, deals with operationalisations of cases where ignorance results in the absence of certain items in a corpus. This question touches on the sign-bound nature of knowledge and is therefore particularly relevant. I discuss it by dealing with a programmatic text by Alan Partington (2014) who asks how one can investigate matters that are absent from corpora. Section 4 gives an overview of studies that explicitly deal with uncertainty and ignorance. I include both corpus linguistic and computational linguistic studies in the literature review. An application sketch (section 5) exemplifies the challenges and opportunities of digital uncertainty research, using an approach we have developed elsewhere (Müller/Bartsch/Zinn 2021). The explanations are concluded by a résumé and outlook (section 6).

Wherever possible, complementary analyses of academic discourse enrich the literature review. These analyses and also the application sketch are based on the Darmstadt International Relations Corpus (DIReC – Müller/Schenk/Steffek 2020), a corpus of political science texts from the field of International Relations. It contains the total number of articles from three leading US-based journals: *International Organization*, *International Security* and *World Politics* from 1974 to 2019, summing up to 5,425 texts and 40,768,562 tokens. For reasons of comparability, I draw on two additional corpora, namely the LOCO corpus and a reference corpus (RC) consisting of randomly selected articles from US national newspapers. LOCO is an 88-million-token corpus composed of topic-matched conspiracy and mainstream documents harvested from 150 websites (Miani/Hills/Bangerter 2022). The presented analyses focus exclusively on the subcorpus representing conspiracy theories (33,506,115 words and 23,937 texts).

2 Ignorance, risk and uncertainty

One of the fundamental distinctions in the literature on ignorance is the one between known and unknown unknowns (Merton 1987, Beck/Wehling 2012, Nielsen/Sørensen 2017: 386 f., Janich 2018: 558, Janich 2020: 50). From a semiotic perspective it is about the sign-bound nature of declarative knowledge: an object identified as 'unknown' or 'uncertain' can be verbalised and linguistically processed just like any other object. Things that are known to be unknown can be described in texts. In academic discourse, in particular, it is common across genres to describe unknown knowledge and to distinguish it from known knowledge. Yet we cannot address something that is reflexively unavailable to us. It is impossible to talk about things whose existence is unknown to the author of a text. Insofar as this unknownness applies

not only to the author but also to everyone else, the absence of the unknown in the text cannot be meaningfully investigated (Warnke 2012). The situation is different, however, with objects that are unknown only to the author of a text, but known to the reader. This can be the case, for instance, if the reader of a historical text reflects a state of knowledge that has changed in the meantime:

> We only come to realize the existence of such unknowns in a retrospect manner, when we become genuinely surprised, for example, in the advent of disasters (Daase/Kessler 2007, Gross 2010). The unexpected occurrence potentially allows us to become aware of our own ignorance and thus may have epistemological value for science, but also moral and social value for society. (Nielsen/Sørensen 2017: 387)

Another case is when the reader belongs to a different domain of knowledge, in the context of which the existence of an unknown fact can be conceptualised. Think, for example, of certain chemical substances that are completely unknown to most people, while experts – through indirect methods of analysis – know that they exist but have not yet been able to identify and describe them. The term *relative ignorance* (Goranko 2021) describes such cases from the reader's perspective as matters of an absence in the text. Under certain circumstances, this absence can be investigated not only linguistically, but also corpus-linguistically (Partington 2014, see below, section 3). Nielsen/Sørensen (2017: 387) also include taboos ("things we don't want to know") and tacit knowledge in the unknown unkowns. Böschen et al. (2010: 784 f.) point out that the various manifestations of the unknown are what establishes non-knowledge research as an independent field within technology assessment – primarily in distinction to research on the concept of risk (Japp 2000: 228). In the latter field, the unknown is treated as a probabilistically calculable quantity and, as it were, conceptually domesticated.

Ignorance research in the line of Merton (1987) treats unknown unknowns as gaps to be filled sooner or later on the path of advancing knowledge. This view is contrasted by critical sociology. Here, ignorance is regarded as a product of diverse discursive practices (e. g. Gross 2010, cf. Nielsen/Sørensen 2017: 386). Against this background, Böschen et al. (2010: 785) diagnose a progressive "'politicization' of ignorance or nonknowledge" since the 1980s. According to the authors, the differentiation of the discursive production not only of knowledge but also of non-knowledge in society and in science, gives rise not only to "cultures of knowledge" (Knorr-Cetina 1999) but also to "cultures of non-knowledge" (Böschen et al. 2010: 787). With this critical focus, Proctor sketches a programme of what he calls "agnotology":

> We need to think about the conscious, unconscious, and structural production of ignorance, its diverse causes and conformations, whether brought about by neglect, forgetfulness, myopia, extinction, secrecy, or suppression. (Proctor 2018: 5)

Müller/Bartsch/Zinn (2021) capture ignorance as one of ten categories of social uncertainty. They measure its occurence using a corpus linguistic annotation approach on a corpus of German and British press articles about the coronavirus pandemic (see below, section 4). Uncertainty, here, is a meta category for various social phenomena that are evidenced in language.[2]

[2] In contrast, Böschen et al. (2010: 808, endnote 1) "emphasize the analytical difference between uncertainty, as a variant of knowledge, if incomplete, and nonknowledge, understood as the absence of knowledge".

This approach examines risk, uncertainty, and ignorance in academic discourse as mediated through national newspaper discourses rather than directly. Ignorance appears as an attribution to science by journalistic, political, and other agents. Müller/Bartsch/Zinn start from the fundamental distinction between personal (or internal) and situational (or external) uncertainty (Kahneman/Tversky 1982):

> While personal uncertainty refers to the representation of cognitive processes, situational uncertainty summarises those forms of uncertainty that are related to an external circumstance or fact. […] we connect social to linguistic research by distinguishing linguistic expressions, which refer to uncertainties as experienced by individuals (such as *I doubt that the sun will shine*), and expressions which refer to the same situation as an objective external condition (*The sun will probably shine tomorrow*). (Müller/Bartsch/Zinn 2021: 500)

Müller/Bartsch/Zinn (2021: 506) distinguish five types of personal uncertainty (anxiety, disagreement, doubt, ignorance, presumption) and five types of situational uncertainty (vagueness, opportunity, possibility, danger, probability). While these types of uncertainty cannot be discussed in detail here, it is important to note that all of them appear in distinctive forms in language (see below, sections 3 and 4).

3 Questions of operationalisation: ignorance and absences in corpora

What are the implications of the introduced terms with regard to a methodology of digital analysis? On a fundamental level, we can only measure in corpora what has been qualified before as a unit of analysis. This limits the analysis to phenomena which are perceptible in an automated way. While known unknowns and uncertainty are expressed in language, unknown unknowns can presumably not be found with methods of analysis that depend on the measurability of data. This does not mean, however, that unknown unknowns cannot be investigated in corpus linguistics. Rather, we need strategies of indirect exploitation of visible language data in research, depending on the type of unknown unknown.

3.1 Categorising and revealing absences

Here, Partington's (2014) notion of "absence" and its exploration in corpora is particularly helpful. In fact, Partington (2014: 122) himself ties in with the non-knowledge debate by making the fundamental distinction between known and unknown absences in corpora. Partington introduces the following categories:

i. known – or suspected, or "searchable" – absence;
ii. unknown absence;
iii. relative absence and absolute absence;
iv. absence from a sizeable corpus;
v. absence from a limited set of texts, including from a specific portion of a corpus;
vi. absence from a position in a single text, including from a location in a phrase;
vii. absence defined as "hidden from open view", that is, hidden meaning. (Partington 2014: 123)

This categorisation is especially interesting for questions of linguistic pragmatics, sociolinguistics, and studies in the line of Foucault's discourse analysis. For our purpose, Partington's mod-

el is highly relevant, but must be considered with caution. First, not all of the referenced absences in texts also correspond to the author's ignorance. For example, the unrealised negation particle *ne* in French sociolects would be considered a known absence in this sense (Partington 2014: 124), although it has obviously nothing to do with ignorance. Second, there is a link to agnotology in Proctor's (2018) critical sense if certain expectable mentions are absent from a corpus or a subset. Partington (2014: 126 f.) gives as example the absence of certain country names in White House press briefings during the Arab Spring. Since it can be assumed that the White House knew about the political situation of the Arab Spring, it is likely that there was at least an attempt to strategically produce ignorance in the public discourse. With this in mind, such absences in corpora are a fundamentally relevant perspective of inquiry. They can be revealed through comparison with other sources of information. Thus, if we want to assume with a critical approach that ignorance is actively produced by suppressing knowledge through semiotic practices, we have to compare different contexts of knowledge constitution with each other. This can be done either by specifically searching for expectable linguistic units – as in the case of the example mentioned – or inductively by linguistic statistical data comparisons between corpora (keyness analysis – Gabrielatos 2018). In this way, depending on the object of investigation and the domain of knowledge, we can investigate not only strategic productions of ignorance such as described above, but also subject matters that are not recognisable and thus knowable in a given (research) discourse.

A procedure mentioned by Partington (2014: 130) is to compare corpora from different time periods and thus to elicit items that were unknown at one time and then became known at a later time. In Partington's example, he compares the complete output of the UK newspapers *The Guardian*, *The Telegraph* und *The Times* over three years (1993, 2005, 2010) by applying keyness analysis.[3] The results reveal not only political and stylistic conjunctures but also the absence of later technical innovations such as *blog*, *website* or *iphone* in 1993. From this absence, ignorance can be inferred by the researcher drawing on their knowledge of the world in interpreting the corpus data (Partington 2014: 130). When we apply this method to DIReC, comparing all IR articles published in the 1970s with the whole corpus, we find leading political theories (*constructivism, contructivist, neorealism, neorealist*) and concepts (*gender, hegemon, globalisation, postcommunist country/state/regime, unipolar world*) to be absent, respectively unknown in 1970s IR discourse – along with other relevant absences, such as those of references to islamism (*Al-Qaida, Taliban, islamist*, *Hamas*). Also, when we measure the negative keywords of all IR articles published in the 2010s as compared to the whole corpus, we find facts that were well known in the earlier discourse and then fell into discursive agnosticism. Examples include forgotten items of concern such as *Antarctica*, *Concorde*, *Euratom*, and *INTELSAT* as well as theoretical frameworks not discussed anymore such as *intergovernalism*, *Marxism-Leninism, neofunctionalism, positivism* and *supranationalism*.[4] In the case of *third world* and related phrases, such as *third-world countries/markets/leaders,* their vanishing is not so much due to agnostics as it reflects the growing awareness towards the exclusionary effects of using this designation.

---

[3] Keyness analysis is a large and dynamic field of research, especially in corpus-based discourse studies. Gabrielatos (2018) and Rayson/Potts (2020) give an overview of new developments. I focus exclusively on Partington's methodological idea that keywords can be used to examine absences in corpora.

[4] All items mentioned here come with frequencies > 100 in the overall corpus and frequencies from 0 to 3 in the 2010s subcorpus.

3.2 Hidden meanings as absences

Partington (2014: 133) also discusses "hidden" meanings as absences in texts. Such implicit meanings are of concern both in linguistic pragmatics and discourse analysis. In our context, implicit meanings are relevant when they can be made intersubjectively plausible by text-analytical procedures, while not being reflexively available to the author. This may be the case with certain presuppositions, topoi, or implications. Partington explains that such implicit meanings are difficult to find automatically and typically require qualitative analysis strategies based on concordances. Furthermore, Partington (2014: 134 f.) draws attention to the problem that implicit meanings are interpretative and may therefore depend on ideological and political preconceptions of the analyst, especially in Critical Discourse Analysis (CDA). In this context, Partington (2014: 133) highlights research on the shifting, unspoken connotations ("semantic prosody") of linguistic expressions. As mentioned, in DIReC, a growing consciousness about the problematic semantic prosody of "third world", "third-world countries" etc. leads to the disappearance of those expressions.

A common grammatical source of absences in texts is passive voice:

> One of the most frequently discussed absences in CDA writings is that mention of a certain participant is missing from a certain position in a discourse, for instance, that a certain passive construction has no agent. It is often inferred that therefore agency is being hidden from the reader. (Partington 2014: 135)

While Partington's particular focus here is on CDA with its critique of political discourse and press language, passive voice is an important and common stylistic device in academic texts as well. Passive constructions have notably been criticised in academic discourse as a means of agent deletion; Billig (2008: 791 f.) has applied this criticism to the original texts of critical linguistics itself. However, passive fulfills important functions in academic texts, such as abstracting from the individual case or periphrasing first person statements. Certainly, there is a fine line between abstraction and the construction of ignorance in the sense introduced, and it is not easy to define it in individual cases. In fact, passive voice in academic discourse often serves to reference actions or effects of actions by unknown or irrelevant actors. The following concordance excerpt from DIReC illustrates this ignorance function of passive constructions using the example of International Relations:[5]

> WP_1986_4_1669 be played again by the victors after a state **has been eliminated** is not enough to guarantee that states will not be eliminated. Consider, for
>
> WP_2007_2_1971 1980s, and by the Serbian army in Kosovo **have been described** as indiscriminate instances of mass violence. This practice has been traced to a combination
>
> IS_1988_3_2188 destruction of the Toksan structure reported that 70 villages **had been submerged**, 800 people were dead or missing, large numbers of domestic animals had been
>
> IO_1975_2_246 now scheduled for completion prior to 1980. It **has recently been announced** that construction will begin in 1980 on two 600 mw facilities and it is anticipated

[5] The bold print in all citations is inserted by me in each case to highlight the constructions of interest (MM).

WP_1997_1_1037 that he helped revive continues unabated. While Netti **has been vindicated**, the form and content of his vindication are full of ironies. The spread

IO_2011_4_704 not give a clear indication of what articles ... **have been suspended**."[38] Studies of the American and European human rights conventions reach similar conclusions.[39]

IS_1979_4_2734 , priority in the employment of conventional ASW forces **has been shifted** from counter-SSBN operations in forward areas such as the Eastern Mediterranean, to reinforcing the

WP_1988_2_1810 theory might examine what happens under conditions in which conflict **is predicted to be greatest** – in cases where the status quo is ambiguous or where

In each of these examples, the passive construction causes the agents to be absent from the actions described. These absences refer partly to the level of the research discourse and partly to the level of the researched facts. In the individual case, it would have to be examined whether the omission of the agents has to do with the fact that they are not known, whether it is a matter of strategic application of agnostic practices in the sense of Proctor (2018), or whether the agents are regarded as irrelevant in the sense of complexity reduction.

4 Corpus approaches to ignorance and uncertainty

Apart from the problems posed by unknown unknowns in corpora discussed above, corpus linguistics proves successful in finding, measuring and contextualising overt linguistic indicators of ignorance and uncertainty. A recurring challenge remains the notorious problem of polysemy when working with expression-based indicators of ignorance and uncertainty (Velldal et al. 2012, Janich 2018: 563 f.). There are various approaches to this. All of them deal with a trade-off between the validity of the data studied and its representativeness: The smaller the data set, the more precisely it can be linguistically classified, the more valid the results. However, a small data set may not be representative of the population of interest. Hence, the goal is to describe as large a corpus as possible as precisely as possible with categories of ignorance and uncertainty. The following section summarises different approaches to solving this problem.

4.1 Pattern-based approaches

A simple, yet promising approach is to rely on lexical resources and search words such as *ignorance, unknown, uncertainty, uncertain* in the corpus of interest. A purely form-based analysis like this remains slightly vague, given that these words appear with different readings. But it is an interesting first source of knowledge, or at least heuristics, for specific questions. Müller/Stegmeier (2019) present a comparative study on the climate change discourse, contrasting British and German news coverage of discursive elaborations on risk and uncertainty in the renewable energies' domain. They analyse the word fields of 'riskiness' in both languages as follows: First, they collect synonyms by evaluating dictionaries and lexical resources as well as by collocation analyses in a corpus of topic-specific words with related meanings. Then, they extract passages where expressions for 'risk' or 'uncertainty' occur in the context of re-

newable energies, measure the significant vocabulary and apply a frame model to categorise and quantify relevant concepts of 'riskiness' comparatively. This approach comes with a loss of lexicological precision, but it makes the conceptual fields in both languages comparable, because it accounts for the different structures in the respective lexical fields. To give an example: the findings from the studies cited above suggest that the English noun *risk* is most frequently (and to an increasing extent) used in phrases such as *at risk* (Zinn 2020), *run a risk* and *take a/ the risk (of)*. To express analogous concepts in German, phrases or word formations are used which are based on the word *Gefahr/danger* (e. g. *gefährdet/at risk, Gefahr laufen/run a risk*) (Müller/Mell 2021: 349 f.).

Collins/Nerlich (2016) use semantic tagging with the corpus analysis tool Wmatrix in British press coverage on climate change in order to measure the word field of uncertainty and to analyse its contextualisation in different knowledge domains. They study how uncertainty is discussed in proximity to the climate change debate. For this purpose, the authors annotate newspaper corpora from different time periods using Wmatrix (Rayson 2009), which – among other features – enables corpus annotation with semantic word domains:

> In this way, the tool is inclusive of a wider field of lexical items that are used to denote for example 'uncertainty', such as 'doubt', 'unclear', 'contentious', 'unsure' etc., as well as all morphological forms of the word 'uncertainty' itself ('uncertain', 'uncertainty', 'uncertainties'). (Collins/Nerlich 2016: 3)

The tool is based on similarity measures of semantic expressions that are not domain-specific. The approach allows the authors to identify the word field around 'uncertainty' in a given corpus on a linguistic-systematic level, but it does not take into account context-specific readings. Also, the approach lacks in that more complex, indirect, or context-specific forms of mentioning ignorance and uncertainty cannot always be found. Uncertainty words are often ambiguous and are specified only in the context of a certain lexical environment. For instance, the word *uncertain* can indicate an external situation (1) as well as a state of mind (2):

(1) The world of crime and, therefore, **the sociological study of crime is** vast and **uncertain**. – BNC baby academic

(2) If **we are uncertain about its value** and apply it mainly to assess its worth then it is research. – BNC baby academic

Vold (2006) studies epistemic modality markers indicating uncertainty, such as *seem, suggest, assume, may, might* in English, French and Norwegian research articles belonging to two different disciplines, linguistics and medicine. Since epistemic modal markers are in many cases polysemous, Vold chooses a manual approach based on a small corpus of a total of 120 articles. Using an exploration corpus consisting of a total of 30 articles, she identifies the 11 most frequent epistemic markers for each of the three languages and two disciplines independently, searches for them in the corpus as a whole and annotates the epistemic senses in the concordances. She then measures frequency differences and applies significance tests on the annotated corpus. The results show that significantly more epistemic hedges are used in the Norwegian and English data than in the French corpus. Disciplinary affiliation and gender, on the other hand, did not produce significant frequency differences in the expression of epistemic modality. Nevertheless, Vold detects differences between disciplines regarding the type of markers used.

4.2 Annotation-based approaches

Kanoksilapatham (2005) adopts Swales' (1990) approach to genre analysis in her study of textual features in biochemical research articles and applies it in an annotation study. She annotates the rhetorical steps in these articles, which Swales called *Moves,* and measures the Inter-Annotator Agreement (see below, section 4.3). To this end, she takes Swales' categories to annotate text segments of her corpus, a body of 65 research articles from the field of biochemistry. The annotations are evaluated both qualitatively and quantitatively. Swales (1990) identifies three basic "Moves" for academic introductions: "establishing a territory (establishing the topic)", "establishing a niche (justifying the present study)", and "occupying a niche (describing the present study)". According to Swales, each Move is comprised of one or more "Steps" (Kanoksilapatham 2005: 271). Steps 1 and 2 of Move 2, "establishing a niche", are especially interesting in our context: "indicating a gap" and "raising a question". Both point to known unknowns in the previous research discourse (Step 1) and in the specific field to be addressed in the paper. Kanoksilapatham describes her results and gives examples:

> Move 2: Preparing for the present study draws scientists' attention to weakness in the existing literature and asserts that a particular research question requires an answer. Unlike Move 1, which is always present, Move 2 was recognized in 40 Introductions or 66.66% of the corpus. The data show that Move 2 has two variations: Step 1: Indicating a gap and Step 2: Raising a question. (Kanoksilapatham 2005: 275)

The examples for Move 2, Step 1, "Indicating a gap", given in Kanoksilapatham (2005: 275), show to what extent this textual practice is connected to addressing ignorance. The teleological model of ignorance is used as a basis. This means that a fact which is to be specified as precisely as possible is treated as a known unknown. It is expected to become a known known as a result of the respective research.

(3) The mechanism of processing the nature, 184nt 6S RNA from its precursor **has not been characterized**. – Kanoksilapatham (2005: 275)

(4) Consequently, how related the serotonin N-acetyltransferase catalytic mechanism will be to that of other superfamily members **is unclear**. – Kanoksilapatham (2005: 275)

Move 2, Step 2, "Raising a question", in contrast, has the function of conceptually circling the known unknown with what is already known and of naming the anchor points from the knowledge already known in the research discourse which are necessary for recognising what is still unknown. This takes the form of questions or functionally equivalent constructions such as hypotheses.

(5) The key (as yet unresolved) questions in analysis of dsRNA-associated PTGS are (1) Why are both strands required in the trigger RNA? and (2) How can dsRNA exert an effect at concentrations that are substantially lower than those of the endogenous target RNA? – Kanoksilapatham (2005: 275)

(6) Is conformational stability a determinant of rebonuclease cytotoxicity? – Kanoksilapatham (2005: 275)

Kanoksilapatham (2005: 275) finds evidence for Step 1, "Indicating a gap", in 38 of the 40 texts she examined, while she surprisingly finds evidence for Step 2, "Raising a question", in only 6 of the texts in her corpus.

In the terminology of Bender/Müller (2020), the Steps described in Kanoksilapatham (2005) are linguistic manifestations of "heuristic textual practices". However, they define the term more broadly. Heuristic textual practices are decision-making routines in academic discourse with which new knowledge is connected to unknown knowledge. This can be done by addressing known unknowns as discussed here so far. But there are also heuristic textual practices that are only indirectly based on ignorance, such as 'defining a term', 'stating a thesis' or 'coining an argument'. Bender/Müller (2020) develop a complex, collaboration-based annotation scheme to annotate heuristic textual practices in 65 introductions of dissertations from 13 different subject areas, thus elaborating text-specific action profiles. They examine their distribution across subject areas and constitute four different action types of scientific introductions.

4.3 Approaches based on machine learning

In Becker/Bender/Müller (2020), these collaboratively created annotations serve as the basis for training a recurrent neural network for classifying heuristic textual practices. Their experiments demonstrate that the annotation categories are robust enough to be recognised by the model, which learns similarities between sentence surfaces represented as vectors. They achieve F1 values[6] that range from 0.75 to 0.92 at different annotation levels. Related studies report F1 values between 0.67 and 0.84 (Becker/Bender/Müller 2020: 450-453). While these values are rated as good or very good in computational linguistics, they still indicate an error rate that is generally unacceptable for (discourse) linguistic studies on an appropriately annotated corpus.

An important factor is the way manual annotations are handled in computational linguistics. They are usually not carried out by experts in the respective field, but exclusively by semi-skilled students or laypersons, who are recruited e. g. via crowd-working platforms and are only instructed via guidelines. The underlying model is that of the "representative language user", who has the same competence as all other members of the language community. The same applies to the metrics used for the Inter-Annotator Agreement (Artstein 2017).

The model of the 'representative language user' has its limitations for use in discourse analyses, since it is precisely here that we are dealing with categories that determine the thinking and actions of communication participants, but which are not reflexively available in everyday communicative life (Müller 2015: 16–27). The studies by Bender/Müller (2020) and Becker/Bender/Müller (2020) confront these limitations with their approach based on collaborative expert annotations. Here, the tagset is repeatedly tested, discussed and incrementally developed, in combination with quality assurance procedures, as is standard in computational linguistics (Bender 2020). The described studies have shown that expert annotation increases the accuracy of neural models. However, the results remain unsatisfactory for linguistic research on pragmatic phenomena such as uncertainty and ignorance.

[6] The F1 value describes the average of the ratio of correctly assigned labels to all assigned labels (precision) and the ratio of all correctly assigned labels to all occurrences of the category (recall).

Szarvas et al. (2012: 339) report on corpora annotated for uncertainty in different domains such as biology, medicine, news media, and encyclopedia. They diagnose "many overlaps but differences as well in the understanding of uncertainty, which is sometimes connected to domain- and genre-specific features of the texts." Szarvas et al. (2012) themselves conduct a study on semantic uncertainty in the three domains of biomedicine, encyclopedia, and newswire: They "consider propositions to which no truth value can be attributed, given the speaker's mental state, as instances of semantic uncertainty" (Szarvas et al. 2012: 336). The authors introduce a unified subcategorisation of semantic uncertainty "as different domain applications can apply different uncertainty categories" (Szarvas et al. 2012: 335). This results in the distinction of four types of uncertainty: epistemic (*It may be raining*), doxastic (*He thinks that the earth is flat*), investigative (*We examined the role of NF-kappa B in protein activation*), conditional (*If it rains, we'll stay in*). Based on this categorisation, they normalise the annotation of three existing corpora and present results with an uncertainty cue recognition model for four fine-grained categories of semantic uncertainty. Like all studies of this type, Szarvas et al. (2012) are faced with the task of working through a multi-dimensional field of research with semantic, syntactic and pragmatic aspects. This includes topics of lexical semantics and text pragmatics in addition to interdisciplinary major categories such as modality and conditionality (Janich 2018: 561–565). In order to reduce complexity, the authors refer to a finite dictionary, i. e. a list of lexical cues, which is then applied to the corpora of interest. Simplification strategies are unavoidable, but come at a cost: In Szarvas et al. (2012) the distinction between epistemic and alethic modality (subjective vs. objective uncertainty) is not made. The blanket classification of conditional sentences as uncertainty markers is debatable: factual conditionals place actuality assertions under conditional reservation. The authors do not explain how this can be linked to a marker of uncertainty.

Gombert/Bartsch (2022) use pre-trained transformer language models for the semantic disambiguation of uncertainty expression in a computational linguistic study on hedging and uncertainty marking in academic discourse. These language models provide vector representations of words that encode not only global distributional properties of a lexical item but also local contextual information on each token:

> These representations promise to allow the contextual disambiguation of words which signal uncertainty only in some contexts, as the respective word vectors should differ from certain to uncertain contexts. (Gombert/Bartsch 2022: 1)

Gombert/Bartsch (2022) compare the performance of two state-of-the-art models, namely RoBERTa-large and DistilRoBERTa-base (Liu et al. 2019b, Sanh et al. 2020). To evaluate the capacity of these models for disambiguating uncertainty markers, they use the above described dataset annotated by Szarvas et al. (2012 – "Szeged uncertainty corpus"), and the corpora from the CoNLL-2010 shared task (Farkas et al. 2010). The latter contains data from the biomedical domain and Wikipedia that comes pre-annotated with another classification approach: Hedges and uncertainty cues are annotated in two different – quite simplistic – classification systems. Both models trained by Gombert/Bartsch (2022) clearly outperform the models used in the previous literature. This demonstrates that disambiguation based on contextual word embeddings is a promising direction in corpus-based uncertainty research. Nevertheless, the models rely on the classifications developed by Szarvas et al. (2012) and Farkas et al. (2010), respectively, and thus inherit the vagueness and conceptual problems that have been observed in these classifications.

4.4 Annotation by query

Müller/Bartsch/Zinn (2021) take a different approach, based on manual annotation, frame-linguistic modelling and corpus-linguistic operationalisation. They combine manual annotation and pattern-based searches in order to find semantic concepts of social uncertainty. They develop the two-level tagset introduced above (section 2) with 'ignorance' understood as a form of 'personal uncertainty'. The annotations are performed collaboratively according to bespoke annotation guidelines and by means of the tool INCEpTION (Eckart de Castilho et al. 2018). They measure Inter-Annotator Agreement and establish a gold standard on a pilot corpus. On the basis of the annotated data, the authors work out which patterns are used in the texts to express uncertainty. For this purpose, they apply the Frame Semantics notation system (Johnson et al. 2003). For example, instead of the verb *to fear*, they search for the semi-specific construction $NP_{[cogniser]}$ fears $NP_{[topic]}$.

Besides the obvious lexical items that serve to denote uncertainty (*possible, risky, unknown* etc.) grammar plays an important role here, e. g. epistemic modality ($NP_{[topic]}$ *could/may* VPinf.) or negated knowledge constructions ($NP_{[cogniser]}$ *doesn't know / needs to know / is unaware of* $NP_{[topic]}$ */ It is not known* $CLsub_{[content]}$ */ No one can say/imagine/foresee* $NP_{[topic]}$). The type of ignorance that can be found with this approach belongs to the category of known unknowns and specified ignorance, as only those are indicated linguistically. In order to apply the gold standard to the entire corpus, Müller/Bartsch/Zinn (2021) draw on an annotation-by-query approach (Eckart de Castilho/Bartsch/Gurevych 2012), in which the gold standard data is translated into CQL (Corpus Query Language; Evert et al. 2020) to enable the most precise searches in terms of precision and recall. As the study focuses on linguistic practices indicating and performing social uncertainty that are typically represented on the sentence level, the authors opt for matching sentences (<s>…</s>) with CQL queries. To give an example, this is the CQL expression used to search 'ignorance' in the data:

> "<s>[]*(([word="n't|not|no|none|nobody"%c][]{0,2}[lemma="aware|know.*|understand|idea|evidence|data"])|[word="ignorance|unaware|unknown"%c]|([word="no|nobody"%c][word="one"]{0,1} [word="can"][word="imagine|foresee|say"]))[]*</s>" returned 2,106 matches in 1,619 different texts (Müller/Bartsch/Zinn 2021: appendix)

Hits based on still ambiguous search expressions are categorised manually: *uncertain* and *uncertainty*, for instance, are polysemous, as shown above. One reading points to 'ignorance' in the classification proposed by Müller/Bartsch/Zinn (2021), which could not be identified by pattern matching. For this reason, the results of a query of all token forms around the word family *uncertain* has to be manually categorized:

> "<s>[]*[lemma="uncertain.*"%c][]*</s>" returned 1,140 matches in 918 different texts, manually categorised as "UK_pers_uncertain_ty_ignorance" (460 hits)

The annotated data are then evaluated with respect to their distributional profile. On this basis Müller/Bartsch/Zinn (2021: 521-525) show that the markers for 'anxiety' and 'possibility' stand out in the UK corpus, while there is a significant increase of 'disagreement' markers during the observation period only in the German news discourse. Instances of 'anxiety' decrease in Germany from the end of January onwards, which is likely to reflect the general confidence in the governmental and medical institutions in Germany during the first phase of

the pandemic. In the UK news coverage on the coronavirus pandemic, markers of 'anxiety' are continuously more frequent,

> with a peak in the last week of February, which was caused by a drop of the stock market. Qualitative analysis shows that there is more reporting on personal experiences and the experience of fear and anxiety, most likely due to historical changes in journalism. The low instantiation of 'disagreement' in the early days of the pandemic might relate to bipartisan support at the beginning of the pandemic, while over time the critique increased significantly in German news coverage (Müller/Bartsch/Zinn 2021: 522).

5 An application sketch

Applying the CQL queries developed in Müller/Bartsch/Zinn (2021) to DIReC, the conspiracy theory subcorpus of LOCO and a reference corpus (RC)[7] I obtained the results given in Figures 1 and 2. They are divided along the main categories of uncertainty described in section 2: situational uncertainty (vagueness, opportunity, possibility, danger, probability) and personal uncertainty (anxiety, disagreement, doubt, ignorance, presumption). These results need to be interpreted with care: The operationalisation and CQL queries have been developed on a UK media corpus, meaning that contextual properties and idiosyncracies of the discourses under review were not considered.[8] Furthermore, CQL queries which needed manual categorisation were not applied. This means that those lexical items which could not be disambiguated by formulating CQL patterns were not considered here. An example is the adjective *uncertain* that points to 'situational danger' as well as to 'personal ignorance' (see above).

The results show meaningful differences in the data sets, some of which are expected, but others, however, come as more of a surprise. For instance, the considerably higher frequency of 'probability' formulations in IR discourse is rather a confirmation of what could be expected assuming that academic discourse operates with the calculability of uncertainty – that is, with the concept of 'probability' (cf. Figure 1). Nevertheless, it is remarkable that the authors of conspiracy theories do not follow this scholarly ductus, but rather evidently conceptualise uncertain outcomes of situations in the interpretive frame of 'danger'. The rather neutral 'possibility' frame is used in the conspiracy theories corpus similarly frequent as in newspaper discourse covered in the reference corpus, while its use in the academic DIReC corpus is significantly less frequent[9]. When expressing situational vagueness, in contrast, the conspiracy theories range in frequency at the level of academic discourse, which is significantly below the frequency of vagueness in everyday journalistic language.[10]

---

7 For brief corpus descriptions and further references see above, section 1.

8 However, the CQL queries do not touch on the difference between British and American English at the level of spelling.

9 Significance tested with Log-Likelihood Ratio LLR = 677.76; $p < 0.0005$.

10 LLR = 1514.69; $p < 0.0005$.

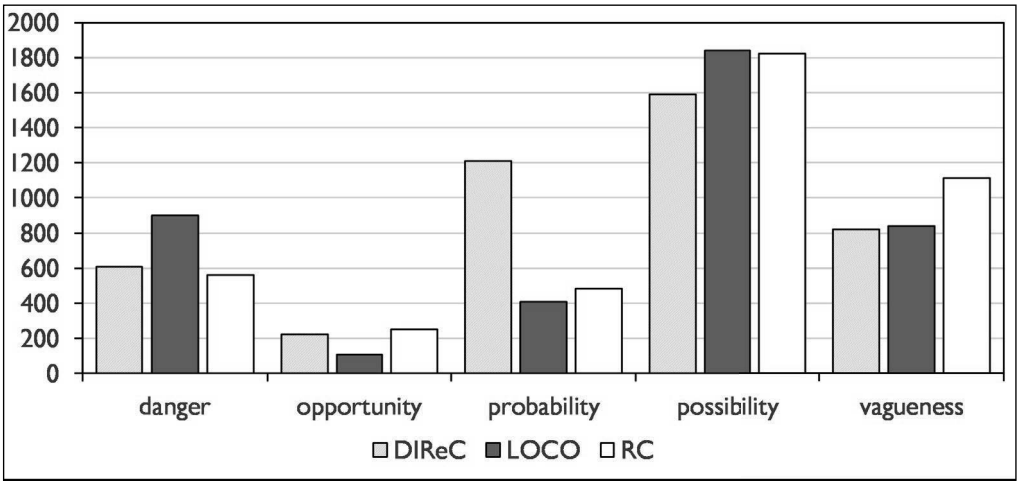

Figure 1: Situational Uncertainty in International Relations journals (DIReC), conspiracy theories (LOCO) and a reference corpus of US newspapers (RC)

On a methodological level, it must be conceded that the word *risk* (noun and verb) is being subsumed here in the 'danger' frame. This is owed to the observation from the original journalistic corpus by Müller/Bartsch/Zinn (2021) that the vast majority of the evidence for *risk* in both English and German is used in the everyday language sense of 'danger', as opposed to the academic sense of 'calculable uncertainty'.[11] Initial preliminary studies show, however, that *risk* is used both in the neutral sense of technology assessment and in the everyday language sense in the academic IR corpus, although the demarcation is not easy in individual cases. Steffek/Müller/Behr (2021) have shown, using the example of *regime*, that IR discourse is characterised precisely by the coexistence of technical terminology and everyday semantics. In contrast, the authors of conspiracy theories use *risk* significantly frequently with quantifiers such as *high, serious, potential, great, low, significant, elevated, severe,* and thus seek a linguistic approximation to the scientific risk discourse. Remarkably, the concept of 'risk' nevertheless has exclusively negative semantics in the sense of 'danger' in an examined sample of 5% of the

[11] For an extensive diachronic study on the semantics of *risk* in the German national parliament cf. Müller/Mell (2021). They examine how the concept of 'risk' and the broader word field of *risk* (including also *threat*, *chance*, *danger*, *hazard*, and *possibility*) have changed over time in parliamentary debate, identifying the most common themes occurring in their co-text using collocation analysis: the vocabulary distribution in the linguistic environment of *risk* is measured and the results compared to a ratio assuming that all words in a data set are equally distributed, which leads to an expected frequency of expressions in the context of *risk*. Müller/Mell (2021) draw on the complete set of German Bundestag minutes to investigate the semantic change of *risk* in German public discourse since 1949. It can be shown that *risk* is developing from a neutral term derived from the technical language in the fields of health, safety and economics to notions of increasingly catastrophic totality. *Risk* takes on the discourse function of *danger* and has almost exclusively negative connotations. This development begins with the critical technology debates of the 1980s and the entry of the Green Party into the German Bundestag, but then spreads across topical contexts as well as parties and the political spectrum.

evidence.[12] If this is taken into account, the difference in frequency of markers of 'risk' between the academic corpus and the conspiracy corpus becomes even greater.

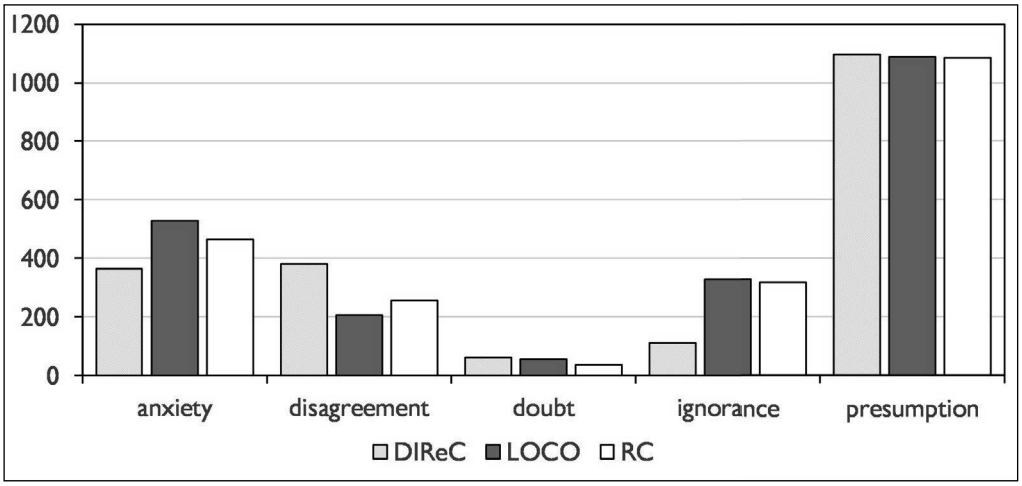

Figure 2: Personal Uncertainty in International Relations journals (DIReC), conspiracy theories (LOCO) and a reference corpus of US newspapers (RC)

To further illustrate the context sensitivity of any operationalisation, we can take the example of 'disagreement'. An important marker for this concept in the source corpus used in Müller/Bartsch/Zinn (2021) was *conflict*. In the IR corpus, however, *conflict* is a keyword for diplomatic and armed disputes between states, which are a central object of the research discourse. This aspect potentially distorts the results regarding the 'disagreement' concept, however, because its usage is aimed precisely at the object dimension of the discourse and does not reveal anything about its mode of interaction. In fact, the value for 'disagreement' with the original search expression is 3.5 times higher in the IR corpus than in LOCO and the reference corpus. But even if we exclude 'conflict' (as documented in Figure 2), the value for disagreement is still almost twice as high as in LOCO, indicating a highly significant difference. Looking at the key vocabulary of the 'disagreement' evidence,[13] we see that *conflict* is indeed by no means the only reason for the high value. Instead, we find numerous indicators of academic dispute such as *whether, question* and *versus* as well as mentions of schools of thought: *neorealist-neoliberal, realist, realism, neorealism*.

The low value for 'ignorance' in the IR corpus may also be surprising, since marking known unknowns is, after all, an integral part of academic practice, as referenced above (see section 2). To explain this, it must be considered that the academic corpus under observation represents a discourse that is not based on a teleology of not-yet-knowledge – as opposed to the life sciences or computer science –, but rather mixes theoretical modelling, political expla-

[12] 398 out of 7,967 occurences were analysed. The mentioned collocates were measured with LLR > 113.

[13] Key lemma list for subcorpus "DISAGREEMENT_DIReC2_wo_conflict" compared to whole "Darmstadt International Relations Corpus (DIReC)"; using log-likelihood statistic, significance cut-off 0.0001% (adjusted LL threshold = 43.31); items must have minimum frequency 5 in list #1 and 5 in list #2. Showing positively key items only.

nation and contemporary historical narrative (Steffek/Müller/Behr 2021). In this multifaceted notion, 'ignorance' is present on various levels, but it is underrepresented. We find ignorance markers in formulations aimed at academic transparency (7) as well as in cases where ignorance is attributed to the referenced actors at the factual level (8).

(7) There is **no evidence** of support from any foreign government. – DIReC WP_1984_1_1771

(8) Rather, the United States **did not understand** its own allies. – DIReC IS_2006_2_2051

More interesting is the case of conspiracy theories. Here, the frequency of 'ignorance' mentions ranges at the level of that in the reference corpus, but their usage differs. In the newspaper texts, 'ignorance' appears in diffuse contexts, often in personalised stories about individuals and referring to everyday life. In contrast, mentions of 'ignorance' in the conspiracy theories have rather clearly delineable functions: Firstly, there are cases where ignorance is attributed to scientists in order to delegitimise them (9). Secondly, there is an action pattern in which ignorance is attributed to certain social groups in order to distinguish them from the group of those who know. This group can be narrowly defined (10), or encompass all of humanity (11). Another function of mentioning not-knowing is documented in citation (12). It is the topos of unexploredness, in which ignorance is identified with danger, or at least such an equation is insinuated. This function is familiar from discourses around emerging technologies such as green genetic engineering.

(9) Gallo claims he has **no idea** how these animal viruses contaminated his lab. – LOCO C023f7 https://humansbefree.com/2014/12/the-secret-origins-of-aids-facts-fallacies-conspiracy-theories.html

(10) Parents are usually **not made aware** of these risks. – LOCO C01314 https://worldtruth.tv/the-lead-vaccine-developer-comes-clean-so-she-can-sleep-at-night-44-girls-are-officially-known-to-have-died-from-these-vaccines/

(11) We **cannot understand** the world until we appreciate that most leaders are traitors and that mankind is victim of a diabolical conspiracy on an unspeakable scale. – LOCO C05238 https://www.savethemales.ca/000546.html

(12) But of course **nobody really knows** what the long-term health effects will be once humans start eating "synthetic proteins" on a massive scale. – LOCO C031c9 https://humansbefree.com/2019/05/after-reading-this-article-about-the-danger-of-gmos-you-will-probably-never-want-to-eat-gm-food-again.html

6 Summary and outlook

In this article, I have presented different approaches to studying ignorance and uncertainty in academic discourse using digital corpora. It turns out that any analysis must take into account two fundamental aspects: the contextual meaning and epistemological implications of each piece of evidence as well as the discursive conditions of the respective academic subject discourses. Both aspects are essential for linguistically describing these phenomena (Janich/Simmerling 2023: 153, 159). However, this is difficult to achieve without downsizing the corpora to a degree where their representativeness is at stake. In contrast, approaches based on

dictionaries and lexical resources can handle large amounts of data, but have only limited explanatory power. This is due to the vagueness and polysemy of the relevant vocabulary and constructions, even if the polysemy problem can be partially solved by pattern matching. Studies on move analysis and heuristic textual practices have shown that investigating ignorance and uncertainty in academic discourse is primarily an indirect process. In fact, we need to take into account precisely those practices in which ignorance is presupposed, implicated or entailed rather than made explicit. In order to make such cases detectable and fruitful for corpus linguistics, we can draw on Partington's (2014) working concept of absence.

The application sketch presented in section 4 is not to be understood as a research contribution, but instead aims at demonstrating the possibilities and limitations of the digital approach using an empirical example on a multi-domain application. The annotation-by-query approach has the advantage that it is applicable over much larger data sets than a purely qualitative annotation approach. At the same time, it allows for efficient control over textual appearances and potential ambiguity via the concordances available at any time. On the downside, the approach, if thoroughly pursued, requires considerable manual re-categorisation and double quality control and is therefore time and money consuming. Current computational linguistic approaches based on contextual word embeddings promise to deliver accurate and reliable results, using a once-categorised dataset, even on large datasets across different domains. These results, too, come with an application threshold: In order to actually work with them in ignorance and uncertainty research, annotation schemes at the cutting edge of interdisciplinary research as well as training datasets annotated by experts are needed. In summary, a combination of the approaches presented here seems to me to yield best results in further research.

Acknowledgements

I am grateful to the editors and anonymous reviewers of FACHSPRACHE for their constructive feedback on an earlier version of the manuscript. I would like to thank Jenni Ellwanger for her help in revising and editing this article. Any remaining errors and inconsistencies are my responsibility.

References

Artstein, Ron (2017): "Inter-Annotator Agreement." *The Handbook of Linguistic Annotation*. Eds. Nancy Ide / James Pustejofsky. Dordrecht: Springer. 297–313.

Beck, Ulrich / Wehling, Peter (2012): "The Politics of Non-Knowing: An Emerging Area of Social and Political Conflict in Reflexive Modernity." *The Politics of Knowing.* Eds. Patrick Baert / Fernando D. Rubio. London / New York: Routledge. 33–57.

Becker, Maria / Bender, Michael / Müller, Marcus (2020): "Classifying Heuristic Textual Practices in Academic Discourse: A Deep Learning Approach to Pragmatics." *International Journal of Corpus Linguistics* 25.4: 426–460. https://doi.org/10.1075/ijcl.19097.bec (16.02.2023).

Bender, Michael (2020): „Annotation als Methode der digitalen Diskurslinguistik." *Diskurse digital. Theorien – Methoden – Fallstudien* 2.1: 1–35. https://doi.org/10.25521/diskurse-digital.2020.140 (23.02.2023).

Bender, Michael / Müller, Marcus (2020): „Heuristische Textpraktiken. Eine kollaborative Annotationsstudie zum akademischen Diskurs." *Zeitschrift für Germanistische Linguistik (ZGL)* 48.02: 1–46. https://doi.org/10.1515/zgl-2020-0001 (23.02.2023).

Billig, Michael (2008): "The Language of Critical Discourse Analysis: the Case of Nominalization." *Discourse & Society* 19: 783–799.

Böschen, Stefan / Kastenhofer, Karen / Rust, Ina / Soentgen, Jens / Wehling, Peter (2010): "Scientific Nonknowledge and Its Political Dynamics: The Cases of Agri-Biotechnology and Mobile Phoning." *Science, Technology & Human Values* 35.6: 783–811.

Collins, Luke / Nerlich, Brigitte (2016): "Uncertainty Discourses in the Context of Climate Change: A Corpus-assisted Analysis of UK National Newspaper Articles." *Communications – the European Journal of Communication Research* 41.3: 291–313. https://doi.org/10.1515/commun-2016-0009 (23.02.2023).

Daase, Christopher / Kessler, Oliver (2007): "Knowns and Unknowns in the 'War on Terror': Uncertainty and the Political Construction of Danger." *Security Dialogue* 38.4: 411–434. https://doi.org/10.1177/0967010607084994 (23.02.2023).

Eckart de Castilho, Richard / Bartsch, Sabine / Gurevych, Iryna (2012): "CSNIPER – Annotation-by-Query for non Canonical Constructions in Large Corpora." *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics. Jesu Island, Korea. 85–90. https://www.aclweb.org/anthology/P12-3015 (23.02.2023).

Eckart de Castilho, Richard / Klie, Jan-Christoph / Kumar, Naveen / Boullosa, Beto / Gurevych, Iryna (2018): "INCEpTION – Corpus-based Data Science from Scratch." *Digital Infrastructures for Research (DI4R), 9–11 October 2018, Lisbon, Portugal.* https://public.ukp.informatik.tu-darmstadt.de/UKP_Webpage/publications/2018/2018_DI4R_INCEpTION-abstract.pdf (28.02.2023).

Evert, Stephanie / The CWB Development Team (2020): *The IMS Corpus Workbench (CWB). CQP Query Language Tutorial. CWB Version.* http://cwb.sourceforge.net/files/CQP_Tutorial.pdf (23.02.2023).

Farkas, Richárd / Vincze, Veronika / Móra, György / Csirik, János / Szarvas, György (2010): "The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text." *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task. Uppsala, Sweden.* Ed. Association for Computational Linguistics. 1–12.

Gabrielatos, Costas (2018): "Keyness Analysis: Nature, Metrics and Techniques." *Corpus approaches to discourse. A critical review*. Eds. Charlotte Taylor / Anna Marchi. Abingdon: Routledge. 225–258.

Gombert, Sebastian / Bartsch, Sabine (2022): *Transformer-based Architectures for the Detection and Disambiguation of Hedges and Semantic Uncertainty*. Darmstadt (unpublished manuscript).

Goranko, Valentin (2021): "On Relative Ignorance." *Filosofiska Notiser* 8.1: 119–140.

Gross, Matthias (2010): *Ignorance and Surprise: Science, Society, and Ecological Design.* Cambridge, MA: MIT Press.

Huang, Yang / Lowe, Henry J. (2007): "A Novel Hybrid Approach to Automated Negation Detection in Clinical Radiology Reports." *Journal of the American Medical Informatics Association* 14.3: 304–311. https://doi.org/10.1197/jamia.M2284 (23.02.2023).

Janich, Nina (2018): „Nichtwissen und Unsicherheit." *Handbuch Text und Gespräch.* Eds. Karin Birkner / Nina Janich. Boston/Berlin: De Gruyter. 555–583.

Janich, Nina (2020): „Wissenschaftliches Nichtwissen in Text und Diskurs – linguistische Perspektiven." *Wissenschaftsreflexion. Interdisziplinäre Grundlagen und ethische Perspektiven*. Eds. Michael Jungert / Andreas Frewer / Erasmus Mayr. Paderborn: Mentis. 45–68. https://doi.org/10.30965/9783957437372 (23.02.2023).

Janich, Nina / Simmerling, Anne (2023): "Linguistics and Ignorance." *Routledge International Handbook of Ignorance Studies* (2nd Ed.). Eds. Matthias Gross / Linsay McGoey. London / New York: Routledge. 150–164.

Japp, Klaus P. (2000): "Distinguishing Non-Knowledge." *The Canadian Journal of Sociology / Cahiers canadiens de sociologie* 25.2: 225–238.

Jean, Pierre-Antoine / Harispe, Sébastien / Ranwez, Sylvie / Bellot, Patrice / Montmain, Jacky (2016): "Uncertainty Detection in Natural Language: A Probabilistic Model." *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics* (WIMS '16). Ed. Association for Computing Machinery, New York, USA, Article 10. 1–10. https://doi.org/10.1145/2912845.2912873 (23.02.2023).

Kahneman, Daniel / Tversky, Amos (1982): "Variants of Uncertainty." *Cognition* 11.2: 143–157.

Kanoksilapatham, Budsaba (2005): "Rhetorical Structure of Biochemistry Research Articles." *English for Specific Purposes* 24: 269–292.

Knorr-Cetina, Karin (1999): *Epistemic Cultures: How the Sciences Make Knowledge.* Cambridge, MA: Harvard University Press.

Liu, Yinhan / Ott, Myle / Goyal, Naman / Du, Jingfei / Joshi, Mandar / Chen, Danqi / Levy, Omer / Lewis, Mike / Zettlemoyer, Luke / Stoyanov, Veselin (2019): "RoBERTa: A Robustly Optimized BERT Pretraining Approach." https://arxiv.org/abs/1907.11692 (23.02.2023).

Merton, Robert K. (1987): "Three Fragments from a Sociologist's Notebooks: Establishing the Phenomenon, Specified Ignorance, and Strategic Research Materials." *Annual Review of Sociology* 13: 1–28. https://doi.org/10.1146/annurev.so.13.080187.000245 (23.02.2023).

Miani, Alessandro / Hills, Thomas / Bangerter, Adrian (2022): "LOCO: The 88-Million-Word Language of Conspiracy Corpus." *Behav Res* 54: 1794–1817. https://doi.org/10.3758/s13428-021-01698-z (23.02.2023).

Müller, Marcus (2015): *Sprachliches Rollenverhalten. Korpuspragmatische Studien zu divergenten Kontextualisierungen in Mündlichkeit und Schriftlichkeit*. Berlin/Boston: De Gruyter.

Müller, Marcus / Bartsch, Sabine / Zinn, Jens O. (2021): "Communicating the Unknown. An Interdisciplinary Annotation Study of Uncertainty in the Coronavirus Pandemic." *International Journal of Corpus Linguistics* 26.4: 498–531. https://doi.org/10.1075/ijcl.21096.mul (23.02.2023).

Müller, Marcus / Mell, Ruth M. (2021): "'Risk' in Political Discourse. A Corpus Approach to Semantic Change in German Bundestag Debates." *International Journal of Risk Research* 25.3: 347–362. https://doi.org/10.1080/13669877.2021.1913631 (23.02.2023).

Müller, Marcus / Schenk, Ana / Steffek, Jens (2020): *The Darmstadt International Relations Corpus (DIReC)*. Darmstadt: TUprints. https://doi.org/10.25534/tuprints-00013063 (23.02.2023).

Müller, Marcus / Stegmeier, Jörn (2019): "Investigating Risk, Uncertainty and Normativity within the Framework of Digital Discourse Analysis. The Example of Future Technologies in Climate Change Discourse." *Researching Risk and Uncertainty – Methodologies, Methods and Research Strategies.* Eds. Anna Olofsson / Jens O. Zinn. Basingstroke: Palgrave. 309–335.

Nielsen, Kristian H. / Sørensen, Mads P. (2017): "How to Take Non-knowledge Seriously, or 'the Unexpected Virtue of Ignorance.'" *Public Understanding of Science* 26.3: 385–392.

Partington, Alan (2014): "Mind the Gaps. The Role of Corpus Linguistics in Researching Absences." *International Journal of Corpus Linguistics* 19.1: 118–146. https://doi.org/10.1075/ijcl.19.1.05par (23.02.2023).

Proctor, Robert N. (2018): "Agnotology. A Missing Term to Describe the Cultural Production of Ignorance (and Its Study)." *Agnotology. The Making and Unmaking of Ignorance.* Eds. Robert N. Proctor / Londa Schiebinger. Stanford: Stanford University Press. 1–36.

Ratcliff, Chelsea L. / Wicke, Rebekah / Harvill, Blue (2022): "Communicating Uncertainty to the Public during the COVID-19 Pandemic: A Scoping Review of the Literature." *Annals of the International Communication Association* 46.4: 260–289. https://doi.org/10.1080/23808985.2022.2085136 (23.02.2023).

Rayson, Paul (2009): *Wmatrix: A Web-based Corpus Processing Environment*. Computing Department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix/ (23.02.2023).

Rayson, Paul / Potts, Amanda (2020): "Analysing Keyword Lists." *A Practical Handbook of Corpus Linguistics*. Eds. Magali Paquot / Stefan Th. Gries. Cham: Springer. 119–139.

Sanh, Victor / Debut, Lysandre / Chaumond, Julien / Wolf, Thomas (2020): *Distilbert, a Distilled Version of Bert: Smaller, Faster, Cheaper and Lighter.* https://arxiv.org/abs/1910.01108 (23.02.2023).

Steffek, Jens / Müller, Marcus / Behr, Hartmut (2021): "Terminological Entrepreneurs and Discursive Shifts in International Relations: How a Discipline Invented the 'International Regime'." *International Studies Review* 23.1: 30–58.

Swales, John (1990): *Genre Analysis*. Cambridge: Cambridge University Press.

Szarvas, György / Vincze, Veronika / Farkas, Richárd / Móra, György / Gurevych, Iryna (2012): "Cross-Genre and Cross-Domain Detection of Semantic Uncertainty." *Computational Linguistics* 38.2: 335–367. https://doi.org/10.1162/COLI_a_00098 (23.02.2023).

Velldal, Erik / Øvrelid, Lilja / Read, Jonathan / Oepen, Stephan (2012): "Speculation and Negation: Rules, Rankers, and the Role of Syntax." *Computational Linguistics* 38.2: 369–410.

Vold, Eva Thue (2006): "Epistemic Modality Markers in Research Articles. A Cross-linguistic and Crossdisciplinary Study." *International Journal of Applied Linguistics* 16.1: 61–87.

Warnke, Ingo H. (2012): „Diskursive Grenzen des Wissens – Sprachwissenschaftliche Bemerkungen zum Nichtwissen als Erfahrungslosigkeit und Unkenntnis." *Nichtwissenskommunikation in den Wissenschaften. Interdisziplinäre Zugänge*. Eds. Nina Janich / Alfred Nordmann / Liselotte Schebek. Frankfurt a. M.: Lang. 51–69.

Zinn, Jens O. (2020): *The UK 'at Risk'. A Corpus Approach to Historical Social Change 1785–2009*. London: Palgrave.

Zinn, Jens O. / Müller, Marcus (2021): "Understanding Discourse and Language of Risk." *International Journal of Risk Research* 25.3: 271–284. https://doi.org/10.1080/13669877.2021.2020883 (23.02.2023).

Prof. Dr. Marcus Müller
Institute of Linguistics and Literary Studies
Technical University Darmstadt
Residenzschloss 1
64283 Darmstadt
Germany
marcus.mueller@tu-darmstadt.de
https://orcid.org/0000-0003-4921-4512