Do Arguments Migrate? Using NLP for Understanding Academia

Jürgen Neyer[1], Sassan Gholiagha[2], Mitja Sienknecht[3]

[1]European New School of Digital Studies (European University Viadrina)
neyer@europa-uni.de
[2]European New School of Digital Studies (European University Viadrina)
gholiagha@europa-uni.de
[3]European New School of Digital Studies (European University Viadrina)
sienknecht@europa-uni.de

Abstract

A crucial question for academia is the relevance of arguments for scientific progress. Are participants in academic debates open to the arguments and insights of other authors, even if they are embedded in competing research paradigms? Or is discursive openness limited to intra-paradigmatic debates? What are the conditions under which arguments are migrating inside and across paradigms? The paper presents the research design and first results from an ongoing research project which uses machine learning (ML) and natural language processing (NLP) to analyse a large corpus that combines thousands of research articles in International Relations (IR) scholarship. The project sets up the most extensive annotated text corpus available for international relations and trains an algorithm to recognise and qualify arguments according to their theoretical origin, supporting evidence and argumentative structure. It relies on an especially designed domain-level category system for the domain-level annotation and a simplified version of Toulmin's argumentation model for the argumentation-level annotation.

Keywords: NLP, ML, International Relations, Arguing, Academia, Politics

1. Introduction

Arguments are central in social science. Arguments are used to make sense of complex data, challenge assumptions, and develop theories. They are often specific to certain theories and help distinguish between competing theories. But while arguments are often assumed to be theory-specific, an open question in science is under what conditions arguments migrate inside and across paradigms. What kind of arguments have a significant probability of changing another's opinion, and to what extent can a systematic connection between reception intensity and specific features of scientific arguments be empirically proven? The paper presents the research design and first findings from a four-year research project to build a social science artificial intelligence (AI) lab for research-based teaching (SKILL).[1] Relying on computational linguistic and visual analysis of the corpus based on machine learning (ML) and natural language processing (NLP), the project aims to demonstrate the importance of arguments and how they are used in scholarly debates from the field of International Relations (IR) and political debates in the global realm. The results of the project and the interfaces and products developed as part of it can then be employed for both research and teaching.

To this end, the paper presents the theoretical foundations of the project (section 2), epistemological reflections (section 3), the data, the model used, and the methodological approach (section 4), as well as first findings in the conclusion (section 5).

2. Theory

Scientific discourse assumes that argumentative quality matters (Zangl and Zürn 1996, Müller 2004, critically Hanrieder 2011). Arguments are assumed to be assessed according to the merits of their scientific quality. Relevant standards include different features depending on scientific theoretical provenance. Positivist epistemologies emphasise the empirical verifiability of claims and the repeatability of lines of evidence (King, Keohane and Verba 1994). Constructivist epistemologies reject this claim and instead emphasise the subjectivity of observation and, thus, the impossibility of objectively testing claims about social facts (Berger and Luckmann 1966; Kratochwil and Ruggie 1986). Therefore, alternative standards of science are emphasised, such as the detailed and plausible reconstruction of meaning with the aim of making them comprehensible and thus understandable (see Jackson 2011 for an overview of different scientific logics for IR).

Regardless of the respective scientific theoretical orientation, theoretical reflections are, in both cases, endowed with additional plausibility when they are supported by empirical evidence. Both perspectives also share the idea that empirical data only become relevant through their explicit integration into a theoretical context. They furthermore both assume that theoretical perspectives gain traction to the degree that they are explained through an explicit exposition of their premises. The idea that quality matters for arguments to be considered seriously also applies to scientific policy advice. When scientists advise policymakers, they usually assume that their

[1] SKILL is funded by the German Ministry for Education and Research, the Brandenburg Ministry for Science and Culture, and the Thuringian Ministry for Science, Research and Art. It is chaired by Bernd Fröhlich, Katrin Girgensohn, Jürgen Neyer, and Benno Stein.

arguments will be considered if they comply with scientific standards.

However, the assumption of a high relevance of argumentation-specific features for their reception by other scientists and policymakers is not undisputed. Receptions within the scientific community are not only influenced by the quality of the arguments presented but also by their integration into established research networks (Risse, Wemheuer-Vogelaar and Havemann 2020) and sometimes even "citation cartels" (Teodorescu and Andrei 2013). Intellectually challenging positions that deviate from the majority opinion are easily ignored if they are not backed by particularly strong arguments and evidence while complying with lower standards is often good enough for arguments that replicate the mainstream. Thomas Kuhn has prominently pointed out that research programs have their own internal logic, selectively receiving content based on whether it fits into dominant paradigms (Kuhn 1962). Despite high formal quality, arguments would be easily ignored if they ignored dominant understandings of problems and solution strategies (paradigms) and followed unorthodox trajectories.

For policy advice, the assumption applies analogously that scientifically sound arguments are only received by policymakers if they can be reconciled with prevailing political calculations, i.e., are politically opportune (Böcher 2022). Luhmann's thesis of different societal functional systems, each with its own language codes and rationality criteria (Luhmann 1984), also suggests that the idea of a search for truth that integrates functional systems and is based on argumentation is at least optimistic: In science, knowledge is generated within the framework of disciplinary concepts and prevailing epistemological interests. It often sits squarely with the logic of politics in which solutions must be negotiated, and compromises will often be based on different values and interests. Science also involves a continuous critique and problematisation of findings, thus inevitably rejecting any conclusive certainty. This irrevocable uncertainty in science is, in turn, difficult to reconcile with the expectation that policymakers are able to make effective decisions that inspire consent and confidence (cf. Böcher 2022).

The tension between the thesis of an argumentation-based dynamic of scientific discourse, on the one hand, and the indications of non-scientific factors influencing the reception of arguments, on the other hand, gives rise to two interrelated questions. First, what is the significance of the quality of a scientific argument for its reception and the change of another's opinion? Second, to what extent can a systematic connection between reception intensity and specific quality features of scientific arguments be empirically proven?

3. Epistemology

The SKILL project addresses these questions by annotating and subsequently analysing a large corpus of academic articles. It develops an algorithm that can recognise and compare patterns of argumentation structures in the corpus. The algorithm may then be used on other corpora, such as debates within the United Nations or other international fora. The project follows an abductive approach, which is based on a combination of ML and NLP. It allows the combination of quantitative and qualitative methods and thus a "methodological twin-move of making *big data thick* and *thick data big*" (Adler-Nissen *et al.* 2021: 1, emphasis in original).

Abductive approaches to pattern recognition have been quite unusual for the social sciences. They have only recently started to gain some attention in the context of large data sets and have only been slowly considered by the social sciences. This immigration into a theory-driven discipline was triggered by the realisation that computer-based methods can unveil social patterns, which have since long been reflected upon but hardly ever been described empirically. The successes of research driven by big data have underlined that individual decisions often reflect broader social patterns rather than individual reflection (Nassehi 2019, Meyer-Schönberger/ Cukier 2013).

Social action is not only shaped by digitalisation but seems to be highly digitally structured and shaped by patterns of rule-compliant action. Pattern recognition procedures thus apply a methodology very much in line with an important logic of social action. The seemingly naive question of "what is?", which has often been rejected as unscientific up to now, moves to the centre in a recognition-oriented approach. Not the testing of hypotheses or the search for merely subjective meaning inherent in understanding-oriented approaches, but the identification, representation and analysis of regular social phenomena – such as arguments – become the goal of the research process.

4. Methodology and Data

This section provides an overview of our methodology, the models used, and the corpus that will be annotated. The section also provides a detailed description of the training process.

4.1. Argumentation model

We use a model of argumentation which builds on NLP methodology, which enables an algorithm to identify and classify arguments. The methodology holds that text can be made machine-readable by annotating individual sentences, i.e. using clearly defined categories to attach meaning to a text.

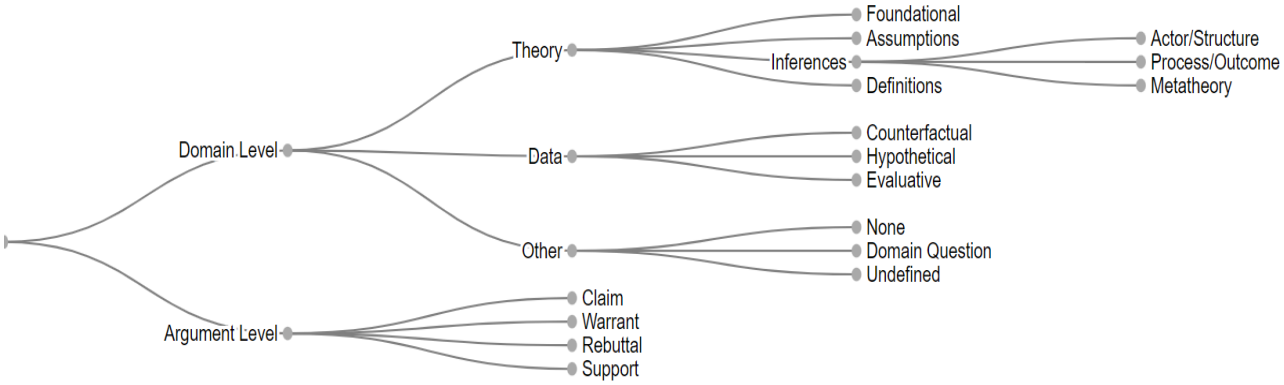

Fig. 1: Overview of categories for annotation. Tree design by Dora Kiesel, Bauhaus University Weimar

The methodology is composed of three main elements:

1. It starts from the assumption that the meaning of sentences can be assessed and understood without referencing the broader context in which they are embedded. Texts are thus decomposed into a set of sentences that are each annotated irrespective of their relationship with provisional or trailing sentences. The decomposition of texts is not unconditional, however. Provisional or trailing sentences are used as an additional resource for annotation if they provide important information without which sentences cannot be properly understood. The process of decomposing texts into sentences is also contextualised by adding relationships between sentences. Sentences that refer to each other and provide an explicit argumentative context are annotated as hanging together. For example, if sentence 1 contains a claim and sentence 2 lists the supporting evidence, then both sentences are annotated as relating to each other.
2. The annotation process works with a category tree that distinguishes between the domain and the argument level (see Figure 1). The domain level refers to propositions which make substantive claims about international politics, such as "war is wrong", "Russia has invaded Ukraine", or the like. In order to allow for a more detailed analysis, the model furthermore distinguishes between theory categories such as foundational, assumption, inferences, and definitions; data categories such as counterfactual, hypothetical, and evaluative; and other categories such as none, domain questions, and undefined. The distinction on the theory part of the category tree between foundational, assumptions, inferences, and definitions allows for considering theoretical contexts such as Realism, Constructivism, etc. Inferences can then be labelled as pertaining to either agents/structure or process/outcome. Additionally, statements could be labelled as providing a metatheoretical assessment of an inference. As a result of this category scheme, a detailed analytical framework emerges that allows the algorithm to search for specific arguments systematically and to relate them to theoretical conceptions.
3. Annotation on the argument level is concerned with the illocutionary aspects of a sentence. Sentences can imply an assertion, support of another claim, a contradiction, or an attack on another position. Annotating these attributes is important for guiding the algorithm in presenting arguments when they are to be set in a discursive context. A Realist debating with a constructivist would, for example, most likely use different concepts on the domain level (emphasising norms rather than interests) and opt on the argument level for a contradiction or an attack to undermine the thrust of a competing argument. Illocutionary annotations are undertaken independently of the material content of a sentence.

4.2. Corpus

The project aims to create and publish an annotated corpus comprising all open-access articles from the most important English-speaking political science journals dealing with international relations.[2] All sentences together will build on a corpus of approximately 800,000 annotated sentences, each with a specific domain meaning and a syntactic (illocutionary) meaning. In this process, subjective meaning

[2] The currently used text corpus comprises a total of 25 different scientific journals with a total of 1980 OpenAccess texts, which are available independently of institutional accesses. These are the American Journal of Political Science, British Journal of Politics and International Relations, Cooperation and Conflict, Ethics & International Affairs, European Journal of International Relations, European Journal of International Security, Foreign Affairs, Global Constitutionalism, Global Society, International Organization, International Security, International Studies Quarterly, International Theory, Journal of Common Market Studies, Journal of Conflict Resolution, Journal of Peace Research, Millennium: Journal of International Studies, Political Research Exchange, Politics and Governance, Politics & Society, Review of International Studies, Security Dialogue, Third World Quarterly, West European Politics, World Politics.

is quasi-reified by being assigned an objectified meaning. An annotated sentence is no longer merely an author's subjective opinion or a recipient's interpretation but becomes a datum with objective domain meaning, syntactic meaning, and a relation to another datum also with objectified domain and syntactic meaning.

This basic sum of annotated sentences represents the raw mass by means of which the algorithm begins to search for specific arguments and patterns of domain and argumentation attributes. With each additional analytic category added to its repertoire, its sensitivity to additional patterns increases, and with each additional text, its ability to process additional statements grows. The resulting dataset allows the algorithm to be trained to identify argumentative patterns from assumptions, processes, and outcomes of different theoretical provenance and to discriminate according to whether and with what kind of structure and evidence they are provided. The result is an instrument that can be used to interrogate texts across theories and time with respect to their argumentative structures and to generate statements about the conditions of their reception or rejection.

4.3. Training

The project invests much effort in the training of the annotators. Here, the training of the annotators (step 1) must be distinguished from training the algorithm (step 2) and from its subsequent independent learning and further data processing (step 3).[3]

The training of the annotators begins with the development of a so-called gold standard. A gold standard is a reference annotation used as a benchmark for annotator performance. In this gold standard, the trainers define a specific mode of annotating sentences with the aim of conveying the underlying principles to the annotators in such a way that they can understand and apply them autonomously. The practice of annotation is trained initially with four central texts characteristic of the four theoretical perspectives of neorealism, liberalism, constructivism and feminism.[4]

These texts are annotated by both student annotators and domain experts (the authors). The annotation process has two aims. First, develop a category system on the domain level that works across different theories, ontologies, and epistemologies (see Figure 1). Second: To train annotators in that category system, refine the category system, and reach a sufficient level of agreement with the gold standard (i.e. the annotation by the senior domain experts) and inter-annotator-agreement. Both are absolutely crucial for step 2.

In step 2, the annotators annotate the large corpus of IR journal articles. Here individual annotators are given different tasks of annotation, with the senior domain experts also annotating some of the corpus. Constant checks of gold standard comparison and inter-annotator reliability ensure sufficient annotation quality. In this step, the algorithm learns to identify arguments relating to theory-specific propositions, to tell, for example, an assumption from an empirical reference and to distinguish between different types of empirical references.

Step 3 of the training grants the algorithm access to the full-text corpus. In this process, the algorithm is set up for (semi-)autonomous annotation and machine learning. It will be closely guided by the annotators and monitored to see if the annotations comply with the standard developed in step 1. This third step leads to a large argumentative repertoire of the algorithm and, thus, significant usability. The repertoire should allow both the systematic search for arguments by users and infer statements about correlations of domain-level features and illocutionary arguments. This approach opens a promising way for answering the research question about the relevance of successful, i.e., persuasive arguments and their domain- and illocutionary features. At a later stage, the algorithm may then be applied to a larger corpus of IR journals or even different corpora, such as debates in the United Nations or the European Union.

This third step leads to a large argumentative repertoire of the algorithm and, thus, significant usability. The repertoire should allow both the systematic search for arguments by users and allow to infer statements about correlations of domain-level features and illocutionary arguments. This opens a promising way for answering the research question about the relevance of successful, i.e., persuasive arguments and their domain- and illocutionary features.

5. Conclusion

The approach taken here to research the relevance of arguments in scientific debates goes a qualitative step further than most previous social science projects. It looks for argumentative patterns in complex communicative acts. Not material reality, but scientific exchange and thus communication about reality is made the object of knowledge. Such a combination of AI/ML and NLP for social scientific reflection and its relevance for political reality has not yet been attempted in this way and to this extent.

Even though SKILL is still in an early phase, first substantial findings can already be reported. The training of the annotators and the implementation of the first annotation exercises on texts from International Relations have underlined the need for, and difficulty of, assigning subjectively meaningful interpretations to an objectifiable schema. This difficulty is first expressed in the definition of separable categories at the domain level. On the one hand, the categories must be specific enough to allow for a high degree of inter-annotator reliability. At the same time, they must be sufficiently general to apply to different theories. What becomes clear in this process is that the structure of

[3] At the time of writing we are in the final stages of step 1.

[4] Kenneth N. Waltz: The Emerging Structure of International Politics, *International Security*. Vol. 18, No. 2 (Fall, 1993), pp. 44-79; Robert D. Putnam, Diplomacy and Domestic Politics: The Logic of Two-Level Games, *International Organization*, Vol. 42, No. 3 (Summer, 1988), pp. 427-460; Finnemore, Martha; Sikkink, Kathryn (1998): International Norm Dynamics and Political Change. In *International Organization* 52 (4), pp. 887–917; Zalewski, Marysia (1995): 'Well, What is the Feminist Perspective on Bosnia?'. In *International Affairs* 71 (2), pp. 339–356.

arguments in scientific texts is far more complex than in other text genres, such as debate articles.

The difficulty of objectifying subjective meanings is also evident in annotators and domain experts working with subjective understandings of IR theories. Establishing an intersubjectively shared understanding thus requires not only mutual explanation but also a high degree of external understanding (Schütze *et al.* 1973). This presents one of the greatest challenges: Is it possible to develop a sufficiently intersubjectively shared understanding of theory without one of the existing interpretations claiming hegemonic status and thus marginalising equally valid interpretations? Or is it the case that the method of pattern recognition by necessity implies the setting of an exclusionary "gold standard"? Are ML and NLP thus necessarily establishing an algorithmic entity with a quasi-scientific "personality" that relies on specific interpretations of reality and will hardly ever be more objective than its annotators?

A final remark relates to the status of theory in a data-driven approach: Social science has, for many years, been dominated by theory. Good scientific work was only too often expected to start with theoretical reflections and use data only to illustrate its findings. Big data, ML and NLP, reverse this methodological bias. The seemingly naive question of "what is?", hitherto often rejected as unscientific, moves to the centre in a pattern-oriented approach. However, an approach to recognising patterns must not be misunderstood as an analytical or theoretical tabula rasa. Unfortunately, exaggerated and misguided misunderstandings of pattern recognition circulate in the literature.

Anderson, for example, fears that in the future digital data analysis will be able to do without researchers since machines could also independently develop the necessary expertise that would be needed in the algorithmic research process (Anderson 2008, Müller and Ritschel 2016: 5). Such fears are based on a misunderstanding of how algorithm-based pattern recognition works. Algorithms can only recognise meaningfully at all, i.e. distinguish relevant from irrelevant, if they have criteria that allow them to make this distinction. For example, an unguided search for patterns may allow the description of reality but will hardly allow any focused statements about scientifically relevant questions. Meaningful recognition, therefore, requires cognition-structuring analytical criteria. These criteria, in turn, cannot be drawn from a conceptual vacuum but must be anchored in theoretical discourses. Like any other social science question, a pattern recognition approach requires a thorough connection to theoretical discourses.

References

Adler-Nissen, R., Eggeling, K.A. and Wangen, P. (2021): Machine Anthropology: A View from International Relations. In *Big Data & Society* 8 (2), 1-6

Anderson, C. (2008): The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Available online at https://www.wired.com/2008/06/pb-theory/

Berger, P.L. and Luckmann, Thomas (1967 [1966]): *The Social Construction of Reality.* First Anchor Books Edition. New York: Anchor Books.

Böcher, Michael (2022): Wie funktioniert wissenschaftliche Politikberatung? in *Forschung und Lehre*, 02.06.2022, https://www.forschung-und-lehre.de/politik/wie-funktioniertwissenschaftliche-politikberatung-475

Hanrieder, Tine (2011): The false promise of the better argument. In *International Theory* 3 (3), 390–415.

Jackson, P. T. (2011): *The conduct of inquiry in international relations. Philosophy of science and its implications for the study of world politics*. London, New York: Routledge

King, G., Keohane, R.O. and Verba, Sidney (1994): *Designing social inquiry*. Princeton, N.J., Chichester: Princeton University Press.

Kuhn, T.S. (1962): *The structure of scientific revolutions*. Chicago: Chicago University Press.

Luhmann, N. (1984): *Soziale Systeme: Grundriss einer allgemeinen Theorie.* Frankfurt am Main: Suhrkamp.

Mayer-Schönberger, V. and Cukier (2013): *Big Data: Die Revolution, die unser Leben verändern wird.* 3rd edition München: Redline Verlag.

Müller, H. (2004): Arguing, Bargaining and All That: Communicative Action, Rationalist Theory and the Logic of Appropriateness in International Relations. In *European Journal of International Relations* 10 (3), 395–435.

Müller, T. and Ritschel, G (2016): Big Data als Theorieersatz? In T. Müller, G. Ritschel, A. Amberger, S. Böschen, R. Broemel, U. Busch et al. (eds.): Big Data als Theorieersatz. *Berliner Debatte Initia*l 4/2016. (2016) 4), 1–8.

Nassehi, A. (2019): *Muster: Theorie der digitalen Gesellschaft.* München: C.H. Beck.

Schütze, F., Meinefeld, W., Springer, W. and Weymann, A. (1973): Grundlagentheoretische Voraussetzungen methodisch kontrollierten Fremdverstehens. In Arbeitsgruppe Bielefelder Soziologen (Hrsg.): *Alltagswissen, Interaktion und gesellschaftliche Wirklichkeit* - Volume 2. Reinbek bei Hamburg: Rowohlt, 433–495.

Teodorescu, D. and Andrei, T. (2014): An examination of "citation circles" for social sciences journals in Eastern European countries. *Scientometrics* 99 (2), 209–231.

Zangl, B.; Zürn, M. (1996): Argumentatives Handeln bei internationalen Verhandlungen: Moderate Anmerkungen zur post-realistischen Debatte. In *Zeitschrift für Internationale Beziehungen* 3, 341-366