

Working paper

The Darmstadt International Relations Corpus (DIREC)

Marcus Müller

Digital Linguistics
Institute of Linguistics and Literary
Studies, Technical University of Darmstadt
marcus.mueller@tu-darmstadt.de

Ana Schenk & Jens Steffek

Transnational Governance
Institute for Political Science, Technical
University of Darmstadt
ana_rabia.schenk@tu-darmstadt.de
steffek@pg.tu-darmstadt.de

Abstract

The Darmstadt International Relations Corpus (DIREC) compiles all articles of the three most influential journals in the field of International Relations – International Organization, World Politics and International Security – from 1974 to 2000. It enables both research in the conceptual history of political science and linguistic analysis of the specialised language. This paper presents the data structure, the user interface and some use cases of DIREC.

Keywords

corpus linguistics · political science · terminology · history of ideas

1. Introduction

In this paper, we introduce the Darmstadt International Relations Corpus (DIREC). DIREC compiles all articles of the three most influential journals in the field of International Relations – International Organization, World Politics and International Security – from 1974 to 2000. It enables both conceptual research in the research debates of that period and linguistic analysis of the specialised language of International Relations.

The purpose of the corpus is to facilitate research on long-term discursive shifts in scholarly research on international relations. It enables us to trace the evolution of the vocabulary that specialists have used to describe and analyse international politics, the global political economy and transnational societal relations. One of the characteristics of IR as a discipline is that many units of analysis escape direct observation or measurement. They need to be imagined. The discipline found its identity through a very peculiar description of the

'international system' in which states interact under conditions of 'anarchy' (Schmidt 1998, Donnelly 2015). That description already entails a definition of the disciplinary agenda and suggested assumptions about the actors and the prevalent rationality of interactions in the system (Wendt 1992, Kratochwil/Ruggie 1986).

In such an imagined social environment, survival is always at stake and social life, to quote Thomas Hobbes, is 'nasty, brutish and short'. In the absence of law-enforcement, violent conflict is always likely among states, and sustained cooperation problematic. States that act under such conditions must accept its imperatives or perish. Vocabularies of IR thus not only invoke a 'global imaginary', which 'makes the global the frame of reference for human action' (Steger 2008: 11). Through their description of what the political world beyond the state is like, they have implications for what is perceived as a possible course of action in the international arena. It is not surprising, therefore, that changes in dominant theoretical orientation among IR scholars are accompanied by changes in vocabulary (Waltz 1979: 12).

DIREC is designed to help researchers interested in international relations theory or the disciplinary language of IR to study these discursive shifts systematically. It is made available via the web-based corpus analysis system CQPWeb and can be accessed on our research platform [Discourse Lab](#) after one-time registration.

2. Corpus Description

2.1. Data and Preprocessing

DIREC is a corpus designed to map the English-language core of academic International Relations, and we selected articles published in three influential American journals to build it. There can be no doubt that since World War II, the core of the academic discipline of IR has been in the United States (Tickner 2013, Holsti 1985, Waever 1998). American universities have more IR degrees on offer than any other country and host the largest IR faculty in the world (Biersteker 2009). Not least, American scholars edit the most influential academic journals in the field. It is unsurprising, therefore, that academic IR has responded to American political debates and foreign policy concerns. It represents a situated worldview. Recent attempts at globalizing the discipline and introducing more diverse perspectives have had limited success only (Hurrell 2016, Knight 2019). In important respects, IR still is an 'American social science' (Hoffmann 1977).

Therefore, we decided to start from the three journals that over decades have counted as the most prestigious publication outlets and set the agenda for the mainstream of the discipline. *International Organization* was founded in 1947. It is the oldest and most important journal in the field, if we follow the citation indices. International Organization covers

the entire field of international affairs. Subject areas include: foreign policies, international relations, international and comparative political economy, security policies, environmental disputes and resolutions, European integration, alliance patterns and war, bargaining and conflict resolution, economic development and adjustment, and international capital movements.

<https://www.cambridge.org/core/journals/international-organization>

World Politics was launched one year later, in 1948. It is open to

research articles that make theoretical and empirical contributions to the literature, and review articles bearing on problems in international relations and comparative politics. The journal does not publish articles on current affairs, policy pieces, or narratives of a journalistic nature. (<https://www.cambridge.org/core/journals/world-politics>)

International Security was established only in 1976. It claims to have “defined the debate on US national security policy and set the agenda for scholarship on international security affairs for more than forty years.” *International Security* publishes

lucid, well-documented essays on the full range of contemporary security issues. Its articles address traditional topics of war and peace, as well as more recent dimensions of security, including environmental, demographic, and humanitarian issues, transnational networks, and emerging technologies. (<https://www.mitpressjournals.org/is>)

All journals are published quarterly. Between 1974 and 2000 this adds up to the following to 2,116 articles and 22,474,553 words in the corpus. Figure 1 shows the distribution over journals:

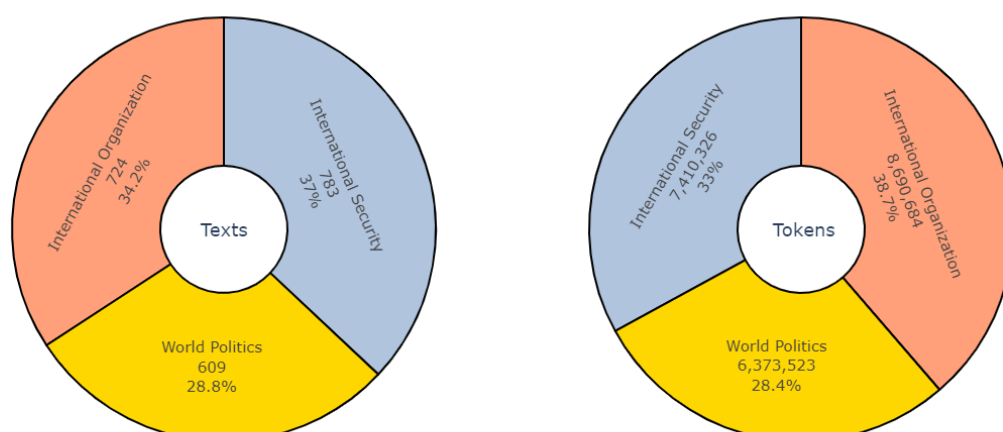


Figure 1: Distributions of articles (texts) and tokens in DReC (chart by Daniel Wachter)

2.1.1. Preprocessing steps

In order to clean up the journal articles, which were originally available as PDF files, and convert them into editable and readable plain text format, we used the optical character recognition (OCR) application ABBYY FineReader. FineReader enables digitizing, retrieving, editing and collaborating on documents in the same workflow.



Figure 2: Abbyy finereader

Automatic recognition of the different text layers made it possible to clean the documents in a targeted manner so that only the main text and copyright notice were retained for further processing. Thus, characters in headers and footers, graphics, marks and smudges of all kinds were manually deleted from the documents to achieve the highest possible precision in the cleanup.

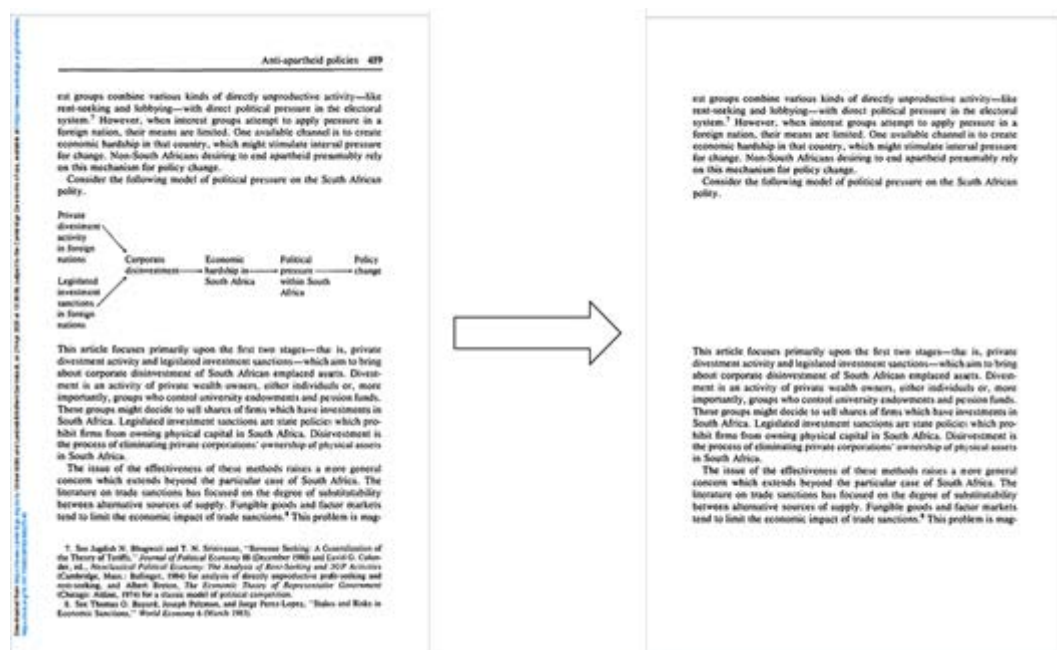


Figure 3: Cleanup: Before and after

Furthermore, footnotes were extracted and transferred to separate text files for each journal issue, so that they could later be added as a footnote subcorpus in CQPweb while remaining related to the associated texts.

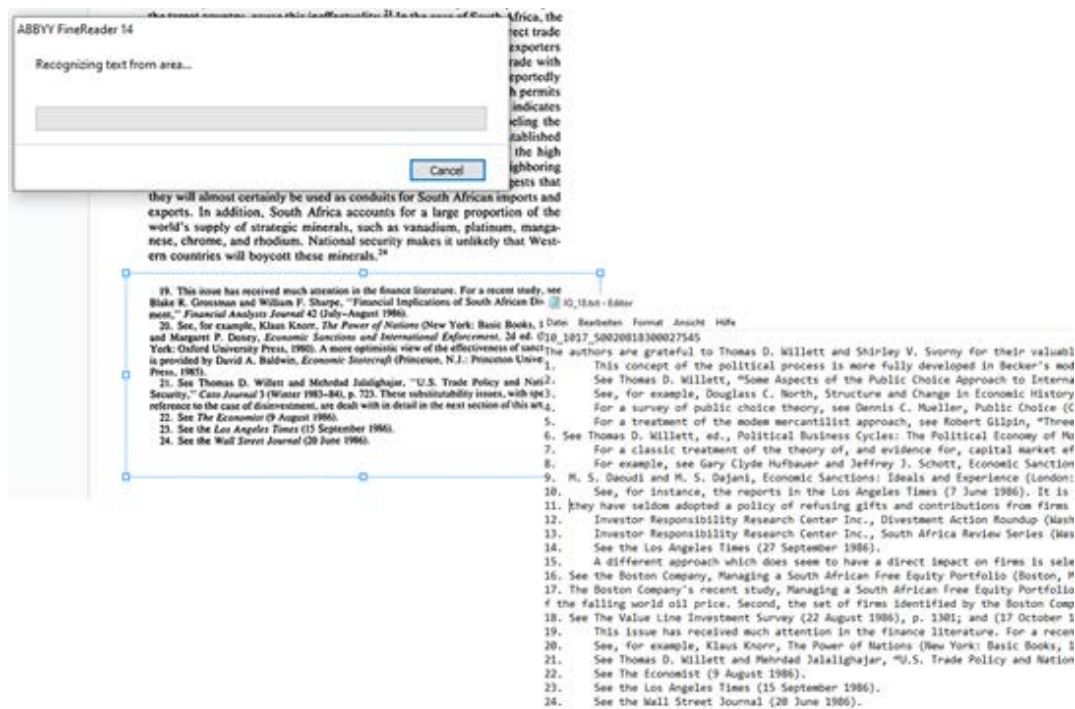


Figure 4: Process of extracting footnotes

The text body and footnote text files resulting from the automatic text recognition process and manual cleaning steps were each named with their respective DOI (Digital Object Identifier) so that the metadata could be accessed and inserted. Most of the further processing steps such as sentence mark-up and ASCII-adjustment were carried out with Python. Text bodies and footnotes were tokenised and tagged. Tokenisation, lemmatisation and part-of-speech-tagging were performed with the TreeTagger (Schmid 1995) using the [BNC tagset](#). The annotated corpora were then imported into the Corpus Workbench and CQPweb for easy access (cf. <http://cwb.sourceforge.net/>, Hardie 2012, Evert & Hardie 2011).

Text and annotations are internally represented in a vertical text format (.vrt format), where each line contains a single text word and tab separated annotations assigned to it. Here is an example of an annotated sentence (taken from Snyder 1991: 121):

```
<s>
It          PNP    it
is          VBZ    be
impossible AJ0    impossible
to          TO0    to
speak      VVI    speak
of          PRF    of
international AJ0    international
relations  NN2    relation
without    PRP    without
referring  VVG    refer
to          PRP    to
alliances  NN2    alliance
;          PUN    ;
the        AT0    the
```

two	CRD	two
often	AV0	often
merge	VVB	merge
in	PRP	in
all	DT0	all
but	CJC	but
name	NN1	name
.	SENT	.

</s>

Annotation makes it possible to search for patterns of language on various levels. If you are, for example, interested in how different forms of crisis are discussed in IR, you might start with searching for the pattern adjective + *crisis*. For addressing the part-of-speech level and the word level within the same query you will need to refer to the Corpus Query Language [CQL](#). Accordingly, our query would be: `[pos="AJO"] [lemma="crisis"]`. This query returns us a list of 2845 matches in 993 texts with 106 different adjectives attributing “crisis” (extract in figure 5), whose most frequent ones are *economic* / *financial* / *international* / *fiscal* / *cuban* / *polish* / *political* / *Balkan* / *severe* / *moroccan* .

Your query "[pos="AJO"] [lemma="crisis"]" returned 2,845 matches in 993 different texts (in 22,474,533 words [2,763 texts]; frequency: 126.59 instances per million words), ordered randomly

No	Filename	Solution 1 to 50	Page 1 / 57
1	IS_1987_3_2231	discussions in the Cabinet Room on the hardest day of the hardest	nuclear crisis so far, I heard no voice raised in anger and no
2	IO_1986_2_637	position in the economic system.11 By 1970, faced with a growing	economic crisis brought on in part by severe weather conditions, the GDR's leaders
3	IS_2000_1_1327	Harvard University, 1992) : 5 Alasdair MacIntyre, "	Epistemological Crises, Dramatic Narrative, and the Philosophy of Science," Monist
4	WP_1997_1_1027	current global ideological environment does is to ensure that responses to a	genuine crisis of capacity will be defensive. Strategies aimed at increasing state capacity
5	WP_1975_2_1482	the Bomb) . . . " 80 But after the	Cuban crisis, the option of " low-level violence " became more and more
6	IO_1996_4_886	) and test those expectations by exploring cases of UN intervention in	interstate crises Previous research on the UN and recurring conflict When the UN
7	WP_1992_4_1889	in keeping law and order. In fact, he used the	economic crisis to rally support for his position, blaming the economic situation on
8	IO_1993_2_722	to raise restrictionist entrance barriers in hard times or periods of global	economic crisis, when both the pressures for protectionism in developed countries and the

Figure 5: adjective + “crisis” in DIReC (extract of the concordance)

The concordance view (as shown here in Fig. 5) gives the possibility to browse search results in a structured way. The linguistic context of the hits can be sorted and expanded so that the use of words and phrases in the text context can be traced. The concordance also allows access to the metadata of the documents.

2.2. Metadata

Not surprisingly, the main issue of crisis debates in IR is thus economy. In order to learn more about economy crisis debates in IR we could refer to the metadata in the corpus. Our database stores information about the author plus date and place of publication for each text in the corpus (fig. 6) as well as information on the text type (text body vs footnote section). This allows us to investigate the distribution of patterns in our corpus.

Metadata for text IO_1992_1_35	
Text identification code	IO_1992_1_35
Structure: "text_author"	Peter M. Haas first
Structure: "text_copyright"	Downloaded from https://www.cambridge.org/core . Universitäts- und Landesbibliothek Darmstadt, on 26 Feb 2018 at 09:05:53, subject to the Cambridge Core terms of use © by The IO Foundation and the Massachusetts Institute of Technology.
Structure: "text_issue"	1
Structure: "text_journal"	International Organization
Structure: "text_journal_short"	IO
Structure: "text_title"	Introduction epistemic communities and international policy coordination
Structure: "text_type"	body
Structure: "text_volume"	46
Structure: "text_year"	1992
Structure: "text_year_issue"	1992_1
No. words in text	11,504

Figure 6: Metadata in CQPWeb

Given that not only the expression *economic crisis* refers to economy but also the frequent attributes *financial*, *fiscal* and *monetary*, we formulate the query [lemma="economic|financial|fiscal|monetary"][[lemma="crisis"]]. This returns 694 matches in 270 different articles. Referring to metadata, we can now investigate whether these expressions are used in footnotes, which may indicate, among other things, that they often appear in titles. (fig. 7). While the vast majority of the usages of our pattern appears in text bodies we see that its distribution in text bodies and footnotes is almost equal relative to the overall amounts of words in each text type. This suggests that the topic is not only assigned or somehow mentioned in the texts, but is regularly the main subject of articles.

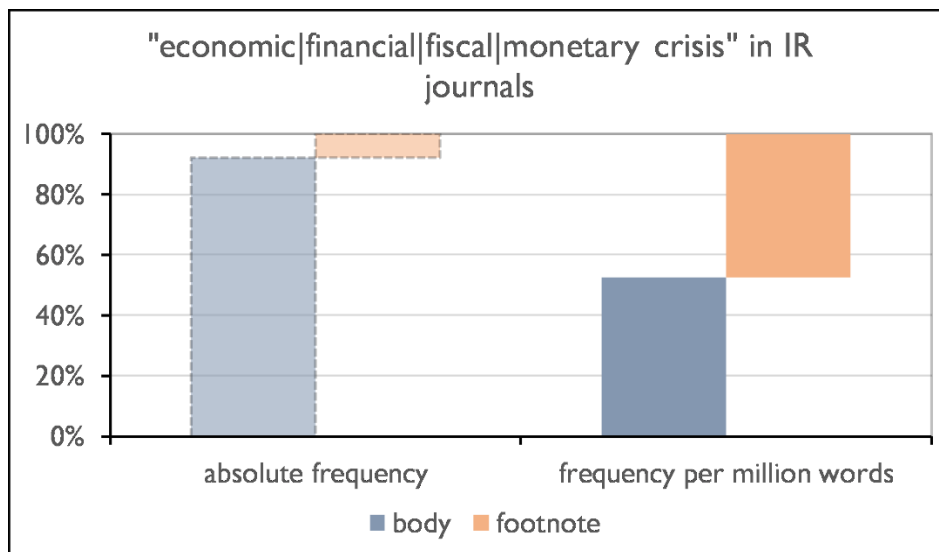


Figure 7: pattern distribution across text types

When exploring the distribution of our pattern across journals, we see clearly that *International Security* is not a platform for debates about economic crises while this topic is discussed primarily in *World Politics* and – significantly less – in *International Organization*.

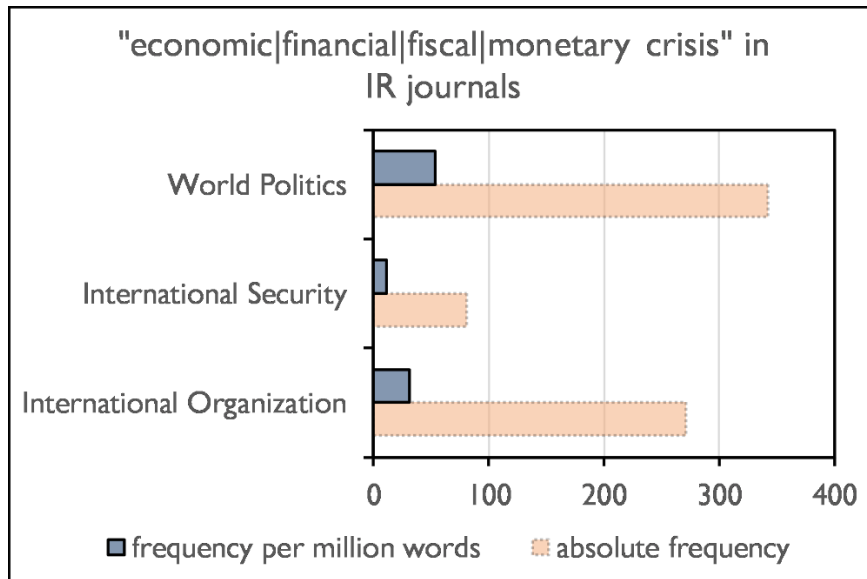


Figure 8: pattern distribution across journals

A distributional exploration across issues shows conjunctures of economic crisis mentions around the years 1983/84, 1994 and in particular 1998-2000. The overall trend is slightly increasing, with considerable variance and peaks between 1985 and 1990. However, note that these peaks indicate single articles that are mainly dealing with economic crisis. The 03/1990 peak, for example, is produced mainly by Marc Linenberg's article on "World Economic Cycles and Central American Political Instability" and Karen L. Remmer's "Democracy and Economic Crisis The Latin American Experience" both published in World Politics 42(03).

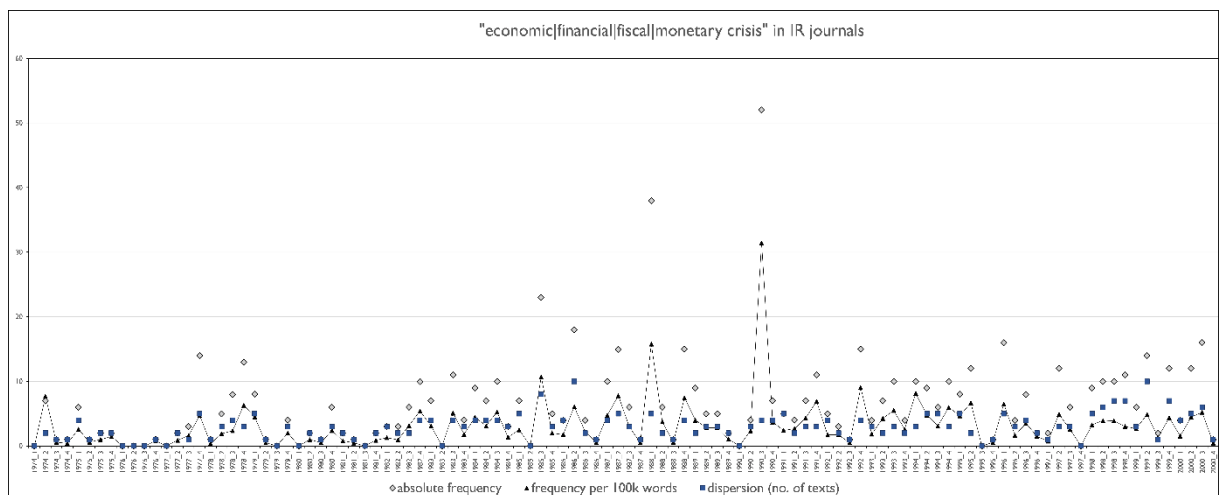


Figure 9: pattern distribution across issues

3. CQPWeb on Discourse Lab

As mentioned above, we make DIReC available via the [CQPWeb analysis system](#), which has been developed by Andrew Hardie at the university of Lancaster. CQPWeb is a graphical user interface based on the [Corpus Query Processor](#) of the IMS Corpus Workbench. It allows both

simply looking up terms, formulations or specific citations in IR discourse and complex queries as indicated above. Search results can also be manually annotated and categorised. For simple word or phrase explorations you just need to enter your query in natural language into the search mask while the more elaborated corpus query language CQL allows more complex queries addressing different levels of annotation. DIReC can hence be used both as a database and as a linguistic corpus.



Figure 10: The CQPWeb graphical user interface

We provide this instance of CQPWeb as part of discourselab.de, a platform for digital discourse analysis particularly designed for interdisciplinary work. Discourse Lab serves to provide resources such as corpora as well as direct collaboration in virtual groups. Discourse Lab also offers comprehensive documentations and tutorials on digital discourse analysis. The research work within Discourse Lab focuses on corpus-supported discourse analysis as well as on method development. It is currently used by 150 international scholars from history, literary studies, linguistics, philosophy, political science, and sociology to carry out interdisciplinary corpus-based discourse research.

4. First results and outlook

The first use case of DIReC was a study on the formation and transformation of the IR term *international regime* (Steffek/Müller/Behr 2020). We traced how the commonplace word *regime* became increasingly used in disciplinary IR in a very specific sense that deviates from everyday usage. We could see that the consensus definition of *international regime* proposed by Stephen Krasner drew on elements already in use since the early 1970s. “Regime” changes from being an ordinary language item in 1974 to a scientific term controlled by a widely accepted definition in 1982. We were able to identify very similar attempts to define “regimes,” which, however, have not been as influential (Müller/Behr/Steffek 2019). Before “regime” became adopted in the discipline in the early 1980s and controlled by an explicit definition, we could observe a period of both competition and academic collaboration, struggling to conceptualize the empirically increasing phenomenon of political cooperation beyond sovereign states. Our evidence thus shows that Krasner’s canonical definition does not fall from the sky but rather develops and grows in discourse.

Our findings suggest studying IR and its theoretical household(s) as assemblages, medleys, and developments of discursive shifts. Drawing from these results we plan to explore the systematicity of terminological development in IR. DIReC will be an important resource for this undertaking enabling large scale linguistic analyses through time.

Acknowledgements

This corpus has been made available in cooperation with [Cambridge University Press](#).

Subject to the Cambridge Core terms of use.

© by The IO Foundation and the Massachusetts Institute of Technology.

We are grateful to Carolina Cardenas, Carina Kiemes, Maxine Schilde, Kateryna Shutiuk, Jörn Stegmeier and Mira Winkelbrandt for their help in compiling DIReC. Daniel Wachter gave us valuable comments on a previous version of this paper and provided us with figure 1.

References

Biersteker, Thomas J. (2009) The Parochialism of Hegemony: Challenges for “American” International Relations. In: Tickner AB and Waever O (eds) International Relations Scholarship Around the World. London: Routledge, 308-327.

Donnelly, Jack. (2015) The discourse of anarchy in IR. *International Theory* 7: 393-425.

Evert, Stefan; Hardie, Andrew (2011): „Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium“. In: Proceedings of the Corpus Linguistics 2011 conference, University of Birmingham, UK. University of Birmingham, UK. Accessible online: <http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-153.pdf>. (last access 20.05.2020)

Hardie, Andrew (2012): CQPweb – combining power, flexibility and usability in a corpus analysis tool. In: *International Journal of Corpus Linguistics* 17:3, 380–409. Accessible online: <http://www.lancaster.ac.uk/staff/hardiea/cqpweb-paper.pdf> [last access: 26.05.2020).

Hoffmann, Stanley (1977) An American Social Science: International Relations. *Daedalus* 106: 41-60.

Holsti, Kal J. (1985) *The Dividing Discipline: Hegemony and Diversity in International Theory*, Boston: Allen & Unwin.

Hurrell, Andrew. (2016) Towards the Global Study of International Relations. *Revista Brasileira de Política Internacional* 59: e008.

Knight, Sarah Cleeland. (2019) Even Today, a Western and Gendered Social Science: Persistent Geographic and Gender Biases in Undergraduate IR Teaching. *International Studies Perspectives* 20: 203-225.

Kratochwil, Friedrich and Ruggie, John G. (1986) International Organization: a State of the Art on an Art of the State. *International Organization* 40: 753-776.

Kristensen, Peter Marcus. (2015) Revisiting the “American Social Science”—Mapping the Geography of International Relations. *International Studies Perspectives* 16: 246-269.

Müller, Marcus, Hartmut Behr & Jens Steffek (2019) ‘The Discursive Formation of Key Terms in International Relations. In: Øivin Andersen, Klaus Schubert & Ingrid Simonnæs (eds.): *New Challenges for Research on Language for Special Purposes*. Berlin: Frank & Timme, S. 105-120.

Schmid, Helmut (1995): „Improvements in Part-of-Speech Tagging with an Application to German.“. In: *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland. Dublin, Ireland.

Schmidt, Brian C. (1998) *The Political Discourse of Anarchy: A Disciplinary History of International Relations*, Albany: State University of New York Press.

Snyder, Glenn H. (1991): Alliances balance and stability. In: *International Organization* 45(1): pp. 121-142.

Steffek, Jens, Marcus Müller und Hartmut Behr (2020) ‘Terminological Entrepreneurs and Discursive Shifts in IR Theory: how a Discipline Invented the “International Regime”’, *International Studies Review*, online, doi: <https://doi.org/10.1093/isr/viaa003>.

Tickner, Arlene B. (2013) ‘Core, Periphery and (Neo)imperialist International Relations. *European Journal of International Relations* 19(3): pp. 627-646.

Waever, Ole (1998) The Sociology of a Not So International Discipline: American and European Developments in International Relations. *International Organization* 52: 687-727.

Waltz, Kenneth N. (1979) *Theory of International Politics*, Reading/Mass.: Addison-Wesley.

Wendt, Alexander. (1992) Anarchy Is What States Make of It: the Social Construction of Power Politics. *International Organization* 46: 391-425.