10. Jg. | H. 2

2
2022

Zeitschrift für Diskursforschung

Journal for Discourse Studies

Sonderausgabe zum zehnjährigen Jubiläum/Special Issue for the Tenth Anniversary

- **Teil A Grundlagen/**
  Part A Fundamentals
- **Teil B Forschungsüberblicke und einzelne Ansätze/**
  Part B Research Overviews and Specific Approaches
- **Teil C Exemplarische Anwendungen/**
  Part C Exemplary Applications
- **Teil D Diagnosen – Diskussionen – Perspektiven/**
  Part D Diagnoses – Discussions – Perspectives

JUVENTA

Inhaltsverzeichnis

Teil C Exemplarische Anwendungen/Part C Exemplary Applications

Teil D Diagnosen – Diskussionen – Perspektiven/Part D Diagnoses – Discussions – Perspectives

Noah Bubenhofer

Corpus Linguistics in Discourse Analysis: No Bodies and no Practices?

Abstract: In corpus linguistics, large amounts of text are analysed using quantitative methods, also for discourse-analytical purposes. However, these methods are limited at first glance because their analyses are restricted to the linguistic surface and do not do justice to the complexity of discourse, and culture. This becomes particularly clear from the perspective of practice theory, which sees body, space and time as constitutive factors of actions. I argue that corpus pragmatic methods can very well integrate the concept of practices into the analysis and show this with an example that uses methods of distributive semantics.

Keywords: discourse, culture, corpus pragmatics, corpus linguistics, distributional semantics

Zusammenfassung: In der Korpuslinguistik werden große Textmengen mit quantitativen Methoden analysiert, auch zu diskursanalytischen Zwecken. Allerdings sind diese Methoden auf den ersten Blick begrenzt, weil ihre Analysen auf die sprachliche Oberfläche beschränkt sind und der Komplexität von Diskurs und Kultur nicht gerecht werden. Dies wird insbesondere aus der Perspektive der Praxistheorie deutlich, die Körper, Raum und Zeit als konstitutive Faktoren von Handlungen versteht. Ich argumentiere, dass korpuspragmatische Methoden das Konzept der Praktiken sehr gut in die Analyse integrieren können und zeige dies an einem Beispiel, das Methoden der distributionellen Semantik verwendet.

Schlagwörter: Diskurse, Kultur, Korpuspragmatik, Korpuslinguistik, distributionelle Semantik

The provocation

When I mentioned the name of Clifford Geertz and his concept of »thick description« in connection with corpus linguistic methods in my post-doctoral lecture at the University of Zurich, resistance arose: A faculty member in the subsequent Q&A session said: Geertz would turn over in his grave if he heard this. Everything I would do methodologically would have nothing whatsoever to do with Geertz. I tried to put forward arguments against it – probably without success.

What in the world made me, a discourse linguist working in corpus linguistics, call Geertz? To the point, this quote:

> »The concept of culture I espouse […] is essentially a semiotic one. Believing, with Max Weber, that man is an animal suspended in webs of significance he himself has spun, I take culture to be those webs, and the analysis of it to be therefore not an experimental science in search of law but an interpretive one in search of meaning.« (Geertz 1973, p. 311)

Corpus linguistics has developed into a sophisticated method within a few decades. Ever larger corpora can be analysed with ever more sophisticated statistical methods. Computational linguistics represents an important source of technological inspiration in this context – especially since the statistical turn, which shows itself e.g. as »distributional semantics«: The combination of huge corpora and neural learning (»deep learning«) leads to so-called language models with which various tasks can be solved very well. Complex meaning relations can be extracted from the data without having to integrate linguistic knowledge into the modelling. A standard evaluation, for example, checks whether the language model can automatically find semantically analogous relations to a given association: Given is the relation »France : Paris« and the counterpart in the relation »Switzerland : ?« is sought. A model trained with the word embedding approach word2vec (Mikolov et al. 2013; for an overview: Lenci 2018) can easily find »Bern« because 1) it has learned the typical contexts of all words in the corpus and 2) it can represent them in a vector space so that 3) analogous semantic relations can be calculated.

This method is interesting for all kinds of applications: machine translation, information extraction from texts, optimisation of searches in large amounts of data, etc. However, interesting applications also arise for research questions from the social sciences and humanities. For example, Kozlowski et al. (2019) showed how this method can be used to extract social stereotypes, for example, about »typical« male and female sports, from the data.

Texts and practices

Nevertheless: methods of distributional semantics are limited to the linguistic surface and take a quantifying approach. This applies to many corpus linguistic methods such as collocation analyses, the calculation of keywords when comparing different corpora, or the analysis of distributions of certain phenomena across the data. How can such quantitative, surface-oriented methods be used for discourse analyses, especially when the goal is to interpret people's self-spun webs of meaning?

There is another complicating aspect: discourse linguistics – especially when it uses corpus linguistic methods – is very text-centered. The theory of practice that has emerged in sociology (Schatzki/Knorr-Cetina/Savigny 2001; Reckwitz 2003) rightly draws attention to this one-sided focus, and calls for understanding body, space, and time as constituting factors of actions. These thoughts have also been received and discussed in linguistics. Thus, Deppermann et al. (2016) propose a praxeological reading of *linguistic* practices and show that language can also be conceptualised as a practice. It is obvious that a phenomenon such as greeting must be understood as a communicative practice in which not only the verbalisation of the greeting is an elementary component but also the physical behaviour and movement in space. The physical behaviour is just as routinised as the linguistic patterns, and the different ways of greeting are socially symbolically loaded. The vocal aspects (intonation curves) also contribute significantly to the constitution

of a communicative practice (vgl. Scharloth 2008 mit dem Beispiel der Fahrkartenkontrolle im Zug). Ultimately, linguistic multimodality research is also strongly influenced by the concept of practices, as is conversation analysis.

However, if one wants to introduce the concept of practices into discourse linguistics, various problems arise. First, the relationship between practices and discourses must be clarified: With Reckwitz (2016, p. 49), one could speak of different analytical strategies, i.e. two different approaches or search and find strategies (ibid., p. 55). Hirschauer and Nübling argue similarly and speak of »layers of meaning of the cultural« as analytical perspectives. They name linguistic structures, discourses, imaginations, and cognitive schemata, situated practices, social structures, and material infrastructures as layers of meaning (Hirschauer/Nübling 2021). Although I consider the concept of language and the associated reduction of linguistic interest to linguistic structures to be clearly too narrow, the analytical separation into the proposed layers of meaning is valuable and helpful in research practice. Furthermore, I think it makes sense to advocate a concept of discourse enriched by practices and to speak from a *linguistic* perspective of a *cultural linguistic* paradigm (Bubenhofer 2009; Bubenhofer/Knuchel/Schüller 2021).

However, I do not want to go further into the relationship between practices and discourses at this point, but rather address a more critical problem: How can a corpus-linguistic discourse and cultural analysis get a look at practices at all?

Corpus pragmatics

To this end, it is first helpful to sharpen the rather heterogeneous concept of corpus linguistics: A useful suggestion is to do this with the term »corpus pragmatics«. This designation, which is now common in the German-speaking world[1] is not used in the same way in the English-speaking world, where both Anglo-Saxon corpus linguistics, with its interest in social discourse, and the French tradition of *léxicometry* and *téxtometry* (Lebart/Salem 1994; transferred to German-speaking research: Scholz/Mattissek 2014; in the context of the software »TXM«: Heiden 2010) have clear points of contact with the understanding of corpus pragmatics.

The term »corpus pragmatics« is intended to make it clear that here corpus linguistics is in the service of a specific research perspective: This perspective is interested in patterns on the linguistic surface, so-called »patterns of language use«, and sees a pragmatic added value inscribed in them. They are understood as effects of social actions, but at the same time have a contextualisation potential, i.e. they also contribute to the emergence of these actions. Feilke understands this view in the context of a »second pragmatic turn« (Feilke 2000), which, following the first pragmatic turn (speech act theory), rejects a systems linguistic reading and makes it clear that linguistic action must be understood in the light of performativity. More modern interactional linguistic theories follow this un-

1 Bubenhofer/Scharloth 2013; Felder/Müller/Vogel 2011b; Meier 2019; Scharloth 2018; Scharloth/Bubenhofer 2011.

derstanding by modelling linguistic interactions as complex situations of language, body, space, and time (Hausendorf 2015; Mondada 2018).

Corpus pragmatic studies aim to investigate this complexity using corpora. Felder et al. (2011a) name society, action, and cognition as cornerstones of an explicative field of corpus pragmatics. They name various fields of linguistics that use corpora to search for typical form-function correlations that stand for the interrelationship between linguistic means and contextual factors: Text linguistics and stylistics, speech act research, sociolinguistics, cognitive research and neurolinguistics, spoken language research and interactional linguistics, and discourse linguistics.

This makes it clear that corpus pragmatic perspectives can certainly take a look at what Geertz called the complex »self-spun web of meaning« of human culture. However, I believe the potential of corpus pragmatic methods has not yet been exhausted. On the one hand, I advocate with Scharloth and Eugster the use of data-driven methods (Scharloth/Eugster/Bubenhofer 2013). On the other hand, I see approaches in distributional semantics whose potential for corpus pragmatics has not yet been exhausted. Finally, I would like to illustrate this with a small example.

Traces of practices

The context of the example is a corpus pragmatic discourse analysis on the COVID-19 discourse in the German-speaking part of Switzerland. The basis of the study is a corpus consisting of three parts: 1) newspaper articles from the five most widely read daily newspapers in German-speaking Switzerland, 2) comments by readers on COVID-19 articles on online news platforms, and 3) a corpus of texts from so-called »alternative media«, Corona measures-sceptical forums and portals of the New Right and followers of conspiracy theories. The corpus comprises over 84 million tokens and covers a period from 1 January 2020 to 31 May 2021 (see Bubenhofer 2022 for details).

The research interest now lies in identifying linguistic patterns that are relevant to the Corona discourse and asking whether they are typical of certain actors and interpreting them. For this purpose, word embeddings were used, and the semantic space of the corpus was calculated with word2vec. This semantic space represents the vocabulary of the corpus (the types) in a vector space based on the typical contexts in which the types appear in the data. The semantic space was automatically trained with the data via Deep Learning and shows which expressions are semantically close to each other because they occur in similar contexts. For example, the following expressions are close to *mask*: *protective mask, face mask, respirator, mouth guard, FFP2 mask, hygienic mask, mask wearing, FFP2 masks, FFP2, wear, muzzle, wearing, no use, cloth mask, mouth, nose protection, surgical masks, FFP 2* etc. (original in German, my translation).

In order to avoid having to determine the set of words of interest for the analysis in a hypothesis-driven way, I applied a clustering procedure to the semantic space: This divides the vocabulary into clusters that contain semantically similar expressions. The vocabulary contains roughly 24,000 types; with 2000 clusters, the average cluster size is

12 types, which is a suitable size to achieve both not too granular but also not too coarse a division of the vocabulary into semantic groups.[2]

The result is a list of clusters with associated key figures: Each type is assigned to a cluster, a label of three types from the cluster describe it. Furthermore, the number of types in the cluster, as well as the frequencies of the types in the data, are given. Table 1 shows an example.

Table 1: Excerpt from list of clusters (basic forms), complete list (originally in German, my translation): https://www.bubenhofer.com/publikationen/corona/

Type	Frequency	Cluster ID	Label (with distance to centroid)	# Types
driver [of the pandemic]	796			
proven	526			
very low	506			
Note thereon	486			
symptomatic	350			
positive effect	296			
Study show	284	425	study prove (0.88983661); scientific prove (0.88387781); prove scientifically (0.88063425)	16
proven	192			
Health risk	192			
negligible	158			
point to	132			
scientific evidence	106			
Study evidence	100			
SARS-CoV-2 virus	94			
international study	81			
scientific evidence	70			

The cluster shown includes all expressions used in connection with scientific studies on the Corona pandemic. This includes the expression ›SARS-CoV-2 virus‹, which is the technical language variant of the various names for the virus.

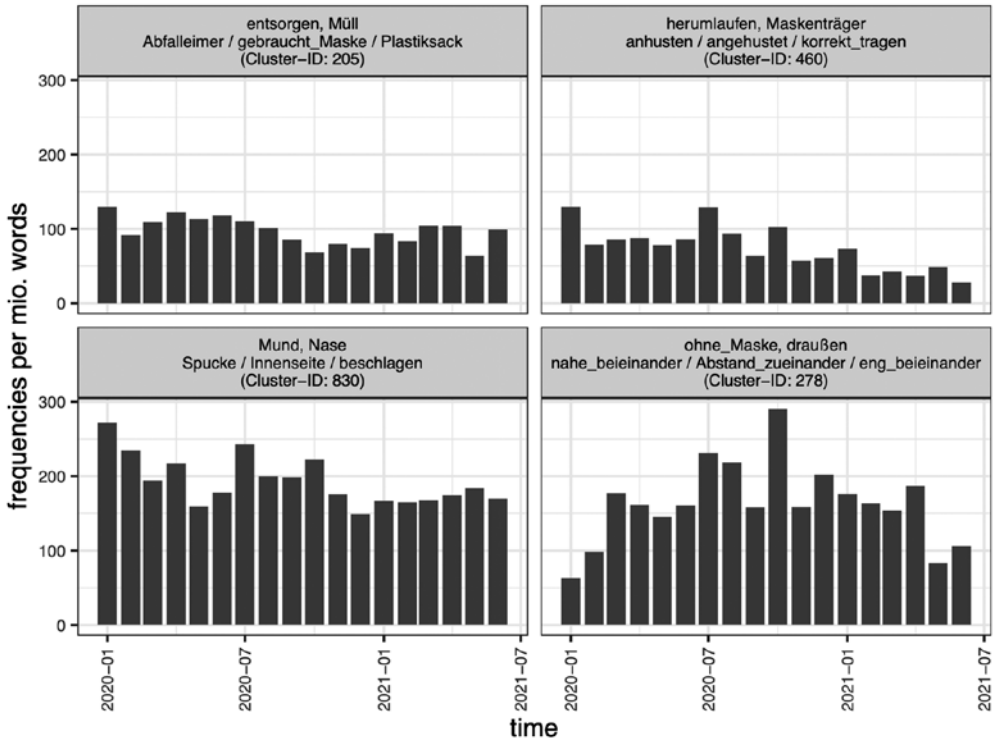
The list of clusters can now be categorised and interpreted according to different research interests. For example, the list contains many clusters of terms denoting the government and indicating different positionings towards it (e.g. ›rule of law‹ vs. ›dictatorship‹ or ›banana republic‹). I cannot discuss this in further detail here (see Bubenhofer

2 The procedure is described precisely in Bubenhofer (2022). The scripts used are available online: https://gitlab.uzh.ch/noah.bubenhofer/distributional-semantics-for-corpus-pragmatics, as are the data: https://www.bubenhofer.com/publikationen/corona/.

2022). However, if one is interested primarily in practices, then several clusters stand out that have to do with thematised or changed (communicative) practices. These include expressions from the field of greeting (*kissing, hugging, greeting, shaking hands, handshake,* etc.) or cleaning (*clean, wash up, cutlery, warm water*).

However, this also includes expressions around wearing masks, as shown in Figures 1 and 2. There are many clusters with different names for masks; more interesting, however, are those that refer to dealing with masks: *without a mask, outside, close to each other, distance from each other, close together* (cluster id 278)*; walking around, mask wearer, coughing on, coughed on, wearing correctly* (cluster id 460)*; mouth, nose, spit, inside, misted* (cluster id 830)*; dispose of, rubbish, bin, used mask, plastic bag* (cluster id 205).[3]

Figure 1: Temporal distribution of four clusters.

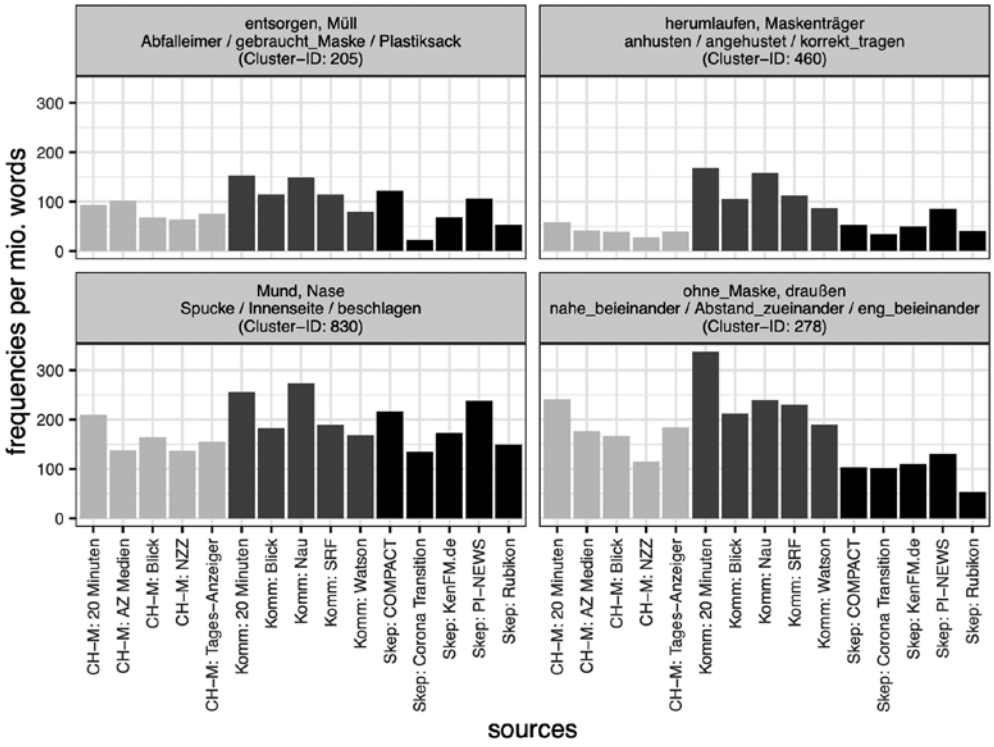


The temporal distribution (Figure 1) and the distribution across the sources (Figure 2) show when and by whom these expressions from these semantic spaces are used. Expressions from cluster 278 (*without mask, outside,* etc.) are used more frequently in July 2020 (introduction of mandatory masks in public transport in Switzerland) and in October

3 See figures for the original German expressions.

2020, for example, while others are more evenly distributed. The distribution across the sources also varies (Figure 2):

Figure 2: Distribution of four clusters across sources: Daily newspapers (light gray), reader comments online news (gray), »alternative media« (black).



Expressions of cluster 460 (*walking around, wearing masks* etc.) are used more frequently in the comments of the readers of the online news platforms than in the other sources. Evidently, this is the place where the correct or incorrect wearing of masks is thematised and discussed.

Conclusion

The brief insight into a corpus pragmatic analysis is intended to show the following: A quantitative, data-driven approach that takes the linguistic surface seriously and does not hastily narrow down to themes can find linguistic sediments of social action. It is apparent that new practices developed or existing ones changed during the Corona pandemic. This is also reflected in linguistic patterns that can be identified and interpreted.

It is crucial that corpus pragmatic methods are understood as reading aids for reading differently–and above all, to be able to read many more texts for linguistic patterns. The methods used must be embedded in a theory that can be connected to discourse and cultural linguistic theories. This is the case with distributional semantics but only if it does not attempt to model »language« in a systems linguistic sense but rather as discursively shaped language usages. The proposed method of clustering and subsequent qualitative categorisation is intended to show how a statistical model can be made interpretable, which is indispensable for a discourse and cultural linguistic perspective.

However, the methods shown need to be further refined: The word embedding method shown (word2vec) is still an unsatisfactory modelling of language use. For example, it does not assume that expressions are mostly polysemous: Each type is represented by *one* vector and therefore occupies a specific place in space. However, language use shows that expressions often have multiple meanings, and accordingly, multiple types must be assumed in the model. Newer word embedding methods take this into account and are, therefore, more plausible (Pelevina et al. 2017). Moreover, semantics is never static but subject to ongoing language change. Recently, there have also been interesting approaches to modelling semantic change (Bianchi 2019).

Moreover, it is worthwhile to embed data-driven approaches in a hermeneutic theory. One possibility, for example, is to adapt grounded theory to corpus pragmatic methods, as Scharloth (2018) showed. As shown above, categories can be derived from data and investigated and differentiated in further analyses. Another example is the embedding of corpus pragmatic analyses in a theory of narration: Based on thousands of women's narratives about the births of their children published in web forums, narrative patterns can be identified in a data-driven way. To do this, however, it is necessary to make such patterns measurable in the first place. The first step is not to understand texts as »bags of words«, but to analyse their sequentiality quantitatively (Bubenhofer 2018, 2020).

I probably did not convince the faculty member I provoked by mentioning Clifford Geertz in a corpus pragmatic context. But I hope that the corpus pragmatic studies will inspire many more experiments.

Acknowledgments: For this publication, use was made of media data made available via Swissdox@LiRI by the Linguistic Research Infrastructure of the University of Zurich (see https://t.uzh.ch/1hI for more information).

Bibliography

Bianchi, F. (2019): Aligning Temporal Word Embeddings with a Compass. In: Medium, https://fedebianchi.medium.com/aligning-temporal-diachronic-word-embeddings-with-a-compass-732ab7427955 (accessed on 27.10.2022).

Bubenhofer, N. (2009): Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse. Berlin und New York: de Gruyter.

Bubenhofer, N. (2018): Serialität der Singularität: Korpusanalyse narrativer Muster in Geburtsberichten. In: Zeitschrift für Literaturwissenschaft und Linguistik 48(2), pp. 357–388.

Bubenhofer, N. (2020): Semantische Äquivalenz in Geburtserzählungen: Anwendung von Word Embeddings. In: Zeitschrift Für Germanistische Linguistik 48(3), pp. 562–89.

Bubenhofer, N. (2022): Exploration Semantischer Räume Im Corona-Diskurs. In: Kämper, H./Plewnia, A. (Hrsg.): Sprache in Politik und Gesellschaft. Berlin und New York: de Gruyter, pp. 197–216.

Bubenhofer, N./Knuchel, D./Schüller, L. (2021): Digitale Kulturlinguistik: Digitalität als Gegenstand und Methode. In: Germanistik in Der Schweiz, Sonderheft Kulturlinguistik 18, pp. 64–96.

Bubenhofer, N./Scharloth, J. (2013): Korpuspragmatische Methoden für Kulturanalytische Fragestellungen. In: Benitt, N./Koch, C./Müller, K./Saage, S./Schüler, L. (Hrsg.): Linguistik als Kulturwissenschaft: Korpus – Kommunikation – Kultur, Giessen Contributions to the Study of Culture. Trier: Wvt Wissenschaftlicher Verlag, pp. 47–66

Deppermann, A./Feilke, H./Linke, A. (2016): Sprachliche und kommunikative Praktiken: Eine Annäherung aus Linguistischer Sicht. In: Deppermann, A./Feilke, H./Linke, A. (Hrsg.): Sprachliche und kommunikative Praktiken. Berlin und Boston: de Gruyter, pp. 1–24.

Feilke, H. (2000): Die pragmatische Wende in der Textlinguistik. In: Brinker, K./Antos, G./Heinemann, W./Sager, S. F. (Hrsg.): Text- und Gesprächslinguistik 1. Halbband. Berlin und New York: de Gruyter, pp. 64–82.

Felder, E./Müller, M./Vogel, F. (2011a): Korpuspragmatik. Paradigma zwischen Handlung, Gesellschaft und Kognition. In: Felder, E./Müller, M./Vogel, F. (Hrsg.): Korpuspragmatik: Thematische Korpora als Basis diskurslinguistischer Analysen. Berlin und New York: de Gruyter, pp. 3–30.

Felder, E./Müller, M./Vogel, F. (Hrsg.) (2011b): Korpuspragmatik: Thematische Korpora als Basis Diskurslinguistischer Analysen. Berlin und New York: de Gruyter.

Geertz, C. (1973): Thick Description: Toward an Interpretative Theory of Culture. In: Geertz, C.: The Interpretation of Cultures. New York: Basic Books, pp. 310–323.

Hausendorf, H. (2015): Interaktionslinguistik. In: Eichinger, L. M. (Hrsg.): Sprachwissenschaft im Fokus. Positionsbestimmungen und Perspektiven. Berlin und Boston: de Gruyter, pp. 43–69.

Heiden, S. (2010): The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In: Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation. Sendai, Japan: Institute of Digital Enhancement of Cognitive Processing, Waseda University, pp. 389–398.

Hirschauer, S./Nübling, D. (2021): Sinnschichten des Kulturellen und die Aggregatszustände der Sprache. In: Dizdar, D./Hirschauer, S./Paulmann, J./Schabacher, G. (Hrsg.): Humandifferenzierung. Disziplinäre Perspektiven und empirische Sondierungen. Weilerswist: Velbrück Wissenschaft, pp. 58–83.

Kozlowski, A. C./Taddy, M./Evans, J. A. (2019): The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. In: *American Sociological Review* 84(5), pp. 905–49.

Lebart, L./Salem, A. (1994): *Statistique textuelle*. Paris: Dunod.

Lenci, A. (2018): Distributional Models of Word Meaning. In: Annual Review of Linguistics 4(1), pp. 151–71.

Meier, S. (2019): Einzelkritiken in der Fußballberichterstattung. Evaluativer Sprachgebrauch aus Korpuspragmatischer Sicht. In: Muttersprache. Vierteljahresschrift Für Deutsche Sprache 129(1), pp. 1–23.

Mikolov, T./Chen, K./Corrado, G./Dean, J. (2013): Efficient Estimation of Word Representations in Vector Space. ArXiv:1301.3781 [Cs].

Mondada, L. (2018): Multiple Temporalities of Language and Body in Interaction: Challenges for Transcribing Multimodality. In: Research on Language and Social Interaction 51(1), pp. 85–106.

Pelevina, M./Arefyev, N./Biemann, C./Panchenko, A. (2017): Making Sense of Word Embeddings. ArXiv:1708.03390.

Reckwitz, A. (2003): Grundelemente einer Theorie sozialer Praktiken: Eine sozialtheoretische Perspektive. In: Zeitschrift Für Soziologie 32(4), pp. 282–301.

Reckwitz, A. (2016): Praktiken und Diskurse: Zur Logik von Praxis-/Diskursformationen. In: Reckwitz, A.: Kreativität und soziale Praxis. Studien zur Sozial- und Gesellschaftstheorie. Bielefeld: transcript, pp. 49–66.

Scharloth, J. (2008): Performanz als Modus des Sprechens und Interaktionsmodalität. Zur linguistischen Fundierung eines kulturwissenschaftlichen Konzeptes. In: Feilke, H./Linke, A. (Hrsg.): Oberfläche und Performanz. Berlin und New York: Max Niemeyer Verlag, pp. 233–254.

Scharloth, J. (2018): Korpuslinguistik für sozial- und kulturanalytische Fragestellungen. Grounded Theory im datengeleiteten Paradigma. In: Kupietz, M./Schmidt, T. (Hrsg.): Korpuslinguistik. Berlin und New York: de Gruyter, pp. 59–78.

Scharloth, J./Bubenhofer, N. (2011): Datengeleitete Korpuspragmatik: Korpusvergleich als Methode der Stilanalyse. In: Felder, E./Müller, M./Vogel, F. (Hrsg.): Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen von Texten und Gesprächen. Berlin und New York: de Gruyter, pp. 195–230.

Scharloth, J./Eugster, D./Bubenhofer, N. (2013): Das Wuchern der Rhizome. Linguistische Diskursanalyse und Data-Driven Turn. In: Busse, D./Teubert, W. (Hrsg.): Linguistische Diskursanalyse. Neue Perspektiven. Wiesbaden: VS, pp. 345–380.

Schatzki, T. R./Knorr-Cetina, K./von Savigny, E. (Hrsg.) (2001): The Practice Turn in Contemporary Theory. New York: Routledge.

Scholz, R./Mattissek, A. (2014): Zwischen Exzellenz und Bildungsstreik. Lexikometrie als Methodik zur Ermittlung semantischer Makrostrukturen des Hochschulreformdiskurses. In: Angermuller, J./Nonhoff, M./Herschinger, E./Macgilchrist, F./Reisigl, M./Wedl, J./Wrana, D./Ziem, A. (Hrsg.): Diskursforschung: Ein interdisziplinäres Handbuch. Bielefeld: transcript, pp. 774–800.

Anschrift:
Noah Bubenhofer
Deutsches Seminar
Schönberggasse 9
CH-8001 Zürich
noah.bubenhofer@ds.uzh.ch
https://www.ds.uzh.ch/p/bubenhofer