herm>A

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Potentials of Automatizing Discourse Analysis

Lessons learned from studying the Phenomenon "Telemedicine"

Paper based on a talk given within the workshop "Digital Traces", University of Bremen

Gertraud Koch and Lina Franken, Project hermA, University of Hamburg

gertraud.koch@uni-hamburg.de, lina.franken@uni-hamburg.de

Version 1.0, 11.12.2018

Within this working paper, we would like to present the first findings of a project on the automatization of discourse analysis. The project is part of the research network hermA (Automated modelling of hermeneutic processes – The use of annotation in social research and the humanities for analyses on health) funded by the Landesforschungsförderung Hamburg, as seed money in the field of digital humanities. It aims for a better understanding of how hermeneutic processes may be enriched or optimized with digital methods (cf. Gaidys et al. 2017). Here, we will present first findings and future work planned from our cultural anthropological subproject. At the same time we want to highlight the relevance of the discussions in the network, with Heike Zinsmeister, Evelyn Gius, Uta Gaidys, Wolfgang Menzel and Dominik Orth, just to mention the Principle Investigators in the network at this point.

We develop our approach on the example of telemedicine in Germany, an emerging phenomenon with various telematics applications in the interaction between doctors and patients since the 1990ies and thus an excellent example for studying social change due to digitalization – which is already all that is to be known about this phenomenon for the following.

2. Discourse Analysis

More relevant in the context of this paper is discourse analysis as our starting point. Discourse analysis in the tradition of the sociology of knowledge (Berger/Luckmann 1969) is a research methodology for gaining an understanding of social orders and how they have emerged over time or are emerging (Keller 2011), like in our example. Social orders are an outcome of discourses, which introduce, negotiate, contest, validate or stigmatize particular stocks of knowledge and ways of knowing, and finally materialize knowledges in regulations, institutions, practices and mind sets – that is at least the wide and largely accepted theory in the tradition of Michel Foucault (as pointed out throughout his work, cf. esp. Foucault 1973). The methodology of discourse analysis, including its intersections to discourse ethnography – which is another methodological discussion – provides a very flexible epistemological toolkit for guiding these studies of knowledge and social orders. We therefore work with a qualitative perspective.

A variety of methods and of materials can be assembled under the umbrella of discourse analysis as research methodology. The variety of social life resonates with the bandwidth of data formats which may be also audio-visual, thus multi-modal data formats matter, even though we concentrate on texts in our project for practical reasons.

An important pillar of the discourse analysis is grounded theory (Glaser/Strauss 1967) as another methodology, which supports the reduction of complexity, one of the most substantial

epistemological processes in all scientific disciplines. Grounded theory is in particular relevant for qualitative research.

Grounded theory procedures, such as theoretical sampling (Glaser 1978), open, axial and selective coding and saturation (Strauss/Corbin 1996), guide the analysis and the selection process of research materials from the collection of the data corpus to analysis of this data (cf. Bryant/Charmaz 2007). The data corpus is built iteratively, not defined initially, because there is little knowledge about the phenome itself and thus about relevant research materials or about explanatory theories. From circle to circle the data corpus and the knowledge grows, thereby facilitates the structuring and narrowing down the research question. Grounded theory is a methodology for discovery in complex social situations and an epistemology for abductive research, in the sense that openness, critical reflexiveness rather than pre-defined theoretical assumptions (Reichertz 2007). In our perception, grounded theory works quite well in balancing the contradictory claims of reflexiveness, openness and the need for narrowing down by structuring social complexity.

Still, in times of digitalization, materials for discourse analyses are available more and more in digital formats, and at the same time in a growing and usually multitudinous number - often called big data (Kinder-Kurlanda 2017). This raises questions about qualitative methodologies: How adequate are qualitative approaches in this situations of big data? On the other side automatization does not look very promising from a qualitative point of view, even though in times of Digital Humanities or Artificial Intelligence. Why?

Automatization works well whenever it comes to large scales, deductive reasoning and the analysis of formal structures. Still, from such approaches usually we do not know so much about what is most relevant in qualitative research: semantics, the variety of meanings, ambiguity. Moreover a lot of resources are needed for automatization: ontologies, tools, capacity for adaptation, corpora etc. If they are not there in the quality needed, automatization of research gets troublesome, which is the case for most qualitative research topics. We study these potentials from a hermeneutic logic and evaluate the Digital Humanities, Computational Social Science and Informatics paradigms from this point of view.

Thinking from the hermeneutic way of thinking, thus crucial questions are: How can the hermeneutic approach be supported by digital methods? And also: When at all should we think about automatizations within a qualitative approach, in terms of efficiency and quality?

This is not a romantic view from hermeneutics in times of paradigm chance towards the digital. Our impression is that there is something to be learned for automatization from the hermeneutic approach. It is first a rational perspective guided by the need of research efficiency and second a reflection about the epistemological implications and the relation of diverse research methodologies.

3. Explorations into Potentials of Automatizing for Discourse Analysis

In the interdisciplinary research project hermA we work from both sides, improving the digital methods, which is the part of computational sciences and computer linguistics, and improving the understanding of hermeneutic processes in respect to including automatization, which is the part of literature studies, health studies and cultural anthropology. Our common starting point is annotation, a very basic epistemological operation, which is done by both humans and computers. For example, humans annotate when they align interpretations to text elements through setting codes, a computer for example when it aligns information about the structure

to a text element, type and token. We will not explore the different forms of coding in discourse analysis because of their semantic complexity, which has been pointed out elsewhere (Strauss/Corbin 1996).

Starting from annotation we explore how computational methods can be used for the diverse annotation processes in discourse analysis. All steps that need to be made in this iterative cycle (cf. Holton 2007; Muckel 2011; Götzö 2014) can be understood as a particular form of annotation: the theoretical sampling for the selection of relevant research materials, the open, axial and selective coding, as well as saturation.

We apply digital methods in our research process as filters of particular quality. Using this understanding of digital methods as filter positions them in a particular place within the research process. They support us to select relevant data, to reduce and structure the complexity given within the diverse materials, which are in our case texts produced on the issue of telemedicine or with relevance to telemedicine. We explored and are still exploring the potential of these filters for

a) finding and selecting relevant expressions in the discourse through crawling / data mining and
b) supporting our coding processes, specifically open coding through structured methods of annotation and text analysis.

We will share the experiences so far. It is work in progress as we emphasized already.

3.1 Approaching Potentials of Automation through Crawling

When looking at the research process of discourse analysis, the first thing to do is to identify relevant expressions and statements within the discourse [Diskursäußerungen] for setting up a corpus of relevant materials. For supporting this step through automatization, text mining is crucial, which is – as Carmen Puchinger (2016) stated in reference to Georg Wiedemann (2016) – not a standard in social science.

Identifying relevant research materials actually is the initial step in a discourse analysis (Glaser/Strauss 1967, 62). We realized that we could not start here with exploring digital methods and potentials for automatization, because we did not know enough about telemedicine for getting started. In particular we were lacking a sound understanding of what telemedicine is, the definition and understanding of the research topic. This is quite usual within the grounded theory approach, since new and emerging phenomenon not known beforehand are studied. And of course for setting up a crawler we need computational skills not to be taken for granted within the humanities.

Thus the start of the discourse analysis was hands on, we first identified who the relevant actors are in the field: physicians, patients, health insurances, IT industry, administration, politics, legislators. Knowing about the strong impact of legislators for ordering social reality we started with analysing their understanding through coding official protocols of debates of the German Federal Parliament (Blaette 2017). What did we learn from this?

With a full-text-search provided on the parliament website, using a first version of a semantic field with keywords describing the phenomenon, we found 120 relevant protocols that we set up as a corpus. What looks like big data, the huge amount of protocols referring to telemedicine turned out to be a very small corpus of relevant statements within this corpus. Often telemedicine is just mentioned as an application example of digital technology in rural

areas or within a discussion on new technologies in general. We retrieved the documents automatically and then had to find out by opening the single documents and reading. As computer linguists we maybe had written a programme, but would this have been efficient?

From the insights in the field gathered with this analysis, we could proceed with searching the web. Thereby, we are again following the theoretical sampling strategy of grounded theory as mentioned above – our corpus is therefore not closed jet and will not be until a quite late time within the research.

A first step is of course searching the web by search engines that we all know and use in everyday life – this is always a good initial step to approach phenomena. Still we face several limitations from a scientific point of view: one cannot go through all search results and does not know about the relevance of a link in the search list.

Even though we started with lists of actors present in the hands-on web search, this soon had to be classified and sorted in order to keep things within an overview. This is what we would have done anyway, but for automatization, this is only the starting point: We clustered our search results without looking at them any further. With this, we produced lists of possible relevant websites that could then be crawled automatically – the deductive part of the work to be done automatically.

We are in the middle of structuring the different ways to handle this. Crawlers can be divided into search engines going deep into a specific website (depth-first) or crawlers that are searching in broader ways (breadth-first), exploring larger parts of the web through the links given. But there are plenty of tools in this area that stay with structured data (like telephone numbers or email addresses), that are not helpful for our approach. We rather need crawlers going in the broader terms in order to find relevant discussions. Furthermore, if we use existing web archives, we need connected snapshot sub-collections (Gossen et al. 2016) since we are interested more in a variety of websites and for a first overview not so much in different versions of websites over time.

As soon as we have relevant texts as starting points, we can use crawler to find data linked to positions identified in advance. With different starting points and therefore different positions within the discourse, we can easily gather different statements on a large scale.

The crawler used so far is IssueCrawler, a tool developed by the digital methods initiative by Richard Roger and his team in Amsterdam (Rogers 2013): it gathers the websites linked on the starting websites and combines them to a network. From this starting point, we can sort our material: With what we know from our hands-on experience, a lexical field of terms is used to index the websites found with the crawler. Other examples of focused crawling, such as the iCrawl tool developed by Risse, Demidova and Gossen (Risse et al. 2014), still have to be examined further.

The potentials for automatization through crawling depend much on the availability of specific resources (indices, lexica, etc.) for the particular knowledge domain. For setting up those tools a lot of interpretative work on the knowledge domain needs to be done and then is invisible part of the algorithms of the tools. When we use generic tools, we do not know much about the decisions in this invisible part. We need a very good understanding on how crawlers work and harvest information, otherwise we hardly know enough about the search principles and thus what parts of the Internet are covered, the part of social reality which is searched. Source criticism as an essential part of social research, relevant in quantitative and qualitative research demands digital tool competence which goes far beyond what we know ourselves and

teach or students. The more sophisticated tools get the more important this competence will be.

3.2. Approaching Potentials for Automation through Co-Reference Analysis

Since we started our discourse analysis with understanding what the phenomenon is about, we started with annotating the protocols of the Bundestag, the Federal German Parliament in this question, we searched for as described.

For analysing the protocols, structuring and narrowing down hypotheses, there are different sorts of annotation, such as open, axial and selective coding, which are applied one after the other, still iteratively (Böhm 2012, 475; Strauss/Corbin 1996, 156f.). The building of categories and structures from these codes, within the hermeneutic circle of analysis, is a manual process by the researcher though supported by QDA-software for organizing and structuring the material that is widely discussed and accepted in qualitative research (Gibbs et al. 2002). The main advantage in our perception is that thus larger amounts of data can be handled in qualitative research.

The automatization of annotation we started to explore were approaches which were available in our team, basically co-reference analysis (Kübler/Zinsmeister 2015, 118; Andresen/Vauth 2018). So we have just started to learn about possible approaches. Since this approach focusses on persons and we are looking at a phenomenon, a lot of additional work and discussions were needed to prepare the tools in a way that they work on texts like the protocols of the parliament.

With co-reference we experience a very indirect way of studying our questions, because we firstly have to understand linguistic concepts and approaches. The annotation is much more fragmented than we would usually do it with annotations in question for our research. And the discussions soon let to very specific questions like whether a law and its draft are co-referent to one another – things that get important if you look at the process of finalizing a law and the discussions surrounding this in parliament. Co-reference is not intended to deal with problems like that, so the annotation took more time than expected and was not to be automatized at all.

What did we learn from that? The search for words in context, topic modelling and other methods of computational analysation thus help us to better understand our data – and the quality of this outcomes can be risen when annotating co-references first. But again, this is only a small amount of new insights gathered (in compare with the data analysed without co-references marked) with a lot of understanding and gaining of tool competences needed first.

For sophisticated methodological approaches in abductive research settings automated annotations need a lot of indirect approaches. In a situation when little is known about a knowledge domain this is not very feasible.

4. Conclusions on the Potentials of Automation for Discourse Analysis and abductive Research

Our input reflects the exploration of existing digital tools into the research process of discourse analysis and how we may learn from them in further ways. Moreover questions about the efficiency of automatizing research processes have to be discussed.

For now, we see the following points for the reflection and realisation of automatization approaches which may be also go beyond qualitative approaches and be useful for social research in a broader sense:

- At the current state of the art, rather than full automation setting automatized filters for reducing and structuring complexity is the aim (and will stay for longer time).
- The potential of automation depends on the research design, in particular on the function theories have for the knowledge production in the research approach, with the highest affinity of deductive research designs to automation and decreasing affinity to inductive and abductive logics of research design.
- Inductive and abductive research approaches apply at particular points in the research process deductive conclusions, here automatization is particularly promising and may go beyond filtering.
- Beyond the deductive, inductive and abductive research design the potentials for automatization of qualitative research depend large and widely on the availability of specific resources (tools, indices, lexica, etc.) for the particular knowledge domain.
- Digital Methods and processes of automation in social analysis are themselves highly depending on interpretative human resources, for example when gold standards of annotated corpora are applied for machine learning. We need a better understanding of how these interpretive processes affect the results of digital tools.
- How are hermeneutic interpretation and automatized analyses are combined it the different approaches of computer scientist and of social researchers? This question needs a lot of further exploration to better understand the implications of either using human or machine filters in research.

References

Andresen, Melanie; Vauth, Michael: Added Value of Coreference Annotation for Character Analysis in Narratives. Proceedings of the Workshop on Annotation in Digital Humanities, 1–6. Sofia, Bulgaria 2018. http://ceur-ws.org/Vol-2155/andresen.pdf

Berger, Peter L.; Luckmann, Thomas: Die gesellschaftliche Konstruktion der Wirklichkeit. Eine Theorie der Wissenssoziologie. Frankfurt a.M. 1969.

Blaette, Andreas (2017): GermaParl. Corpus of Plenary Protocols of the German Bundestag. TEI files, availables at: https://github.com/PolMine/GermaParlTEI.

Bryant, Antony; Charmaz, Kathy (Hg.): The SAGE Handbook of Grounded Theory. Los Angeles 2007.

Foucault, Michel: Archäologie des Wissens. Frankfurt a.M. 1973.

Gaidys, Uta; Gius, Evelyn; Jarchow, Margarete; Koch, Gertraud; Menzel, Wolfgang; Orth, Dominik; Zinsmeister, Heike: Project Descripition. hermA: Automated Modelling of Hermeneutic Processes. In: Hamburger Journal für Kulturanthropologie 7 (2017), S. 119–123.

Gibbs, Graham R.; Friese, Susanne; Mangabeira, Wilma C.: Technikeinsatz im qualitativen Forschungsprozess. Einführung zu FQS Band 3(2). In: Forum Qualitative Sozialforschung 3 (2002), S.

Glaser, Barney G.: Theoretical Sensitivity. Advances in the Methodology of Grounded Theory. Mill Valley, Calif. 1978.

Glaser, Barney G.; Strauss, Anselm L.: Grounded Theory. Strategien qualitativer Forschung. Bern 2010 [1967].

Gossen, Gerhard; Demidova, Elena; Risse, Thomas: Analyzing Web Archives through Topic and Event Focused Sub-Collections. In: Proceedings of Web Science WebSci 2016. https://www.l3s.de/~gossen/publications/gossen_et_al_websci_2016.pdf

Götzö, Monika: Theoriebildung nach Grounded Theory. In: Bischoff, Christine; Oehme-Jüngling, Karoline; Leimgruber, Walter (Hg.): Methoden der Kulturanthropologie. Bern 2014, S. 444–458.

Holton, Judith A.: The Coding Process and Its Challenges. In: Bryant, Antony; Charmaz, Kathy (Hg.): The SAGE Handbook of Grounded Theory. Los Angeles 2007, S. 265–289.

Keller, Reiner: Wissenssoziologische Diskursanalyse. Grundlegung eines Forschungsprogramms. 3. Auflage Wiesbaden 2011 [2005].

Kinder-Kurlanda, Katharina E.: Big Data. In: Koch, Gertraud (Hg.): Digitalisierung. Theorien und Konzepte für die empirische Kulturforschung. Konstanz/München 2017, S. 217–240.

Kübler, Sandra; Zinsmeister, Heike: Corpus Linguistics and Linguistically Annotated Corpora. London/New York 2015.

Muckel, Petra: Die Entwicklung von Kategorien mit der Methode der Grounded Theory. In: Mey, G. Nter; Mruck, Katja (Hg.): Grounded Theory Reader. Wiesbaden 2011, S. 333–352.

Puchinger, Carmen: Die Anwendung von Text Mining in den Sozialwissenschaften. Ein Überblick zum aktuellen Stand der Methode. In: Lemke, Matthias; Wiedemann, Gregor (Hg.): Text Mining in den Sozialwissenschaften. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse. Wiesbaden 2016, S. 117–138.

Reichertz, Jo: Abduction. The Logic of Discovery of Grounded Theory. In: Bryant, Antony; Charmaz, Kathy (Hg.): The SAGE Handbook of Grounded Theory. Los Angeles 2007, S. 214–228.

Risse, Thomas; Demidova, Elena; Gossen, Gerhard: What Do You Want to Collect from the Web? In: Procedings of the Building Web Observatories Workshop BWOW 2014 (2014), S. 1–7.

Rogers, Richard: Digital Methods. 2013.

Strauss, Anselm L.; Corbin, Juliet M.: Grounded Theory. Grundlagen qualitativer Sozialforschung. Weinheim 1996.

Wiedemann, Gregor: Text Mining for Qualitative Data Analysis in the Social Sciences. A Study on Democratic Discourse in Germany. Wiesbaden 2016.