SAMPLING PROCESS-GENERATED DATA: THE CASE OF NEWSPAPERS

Nina Baur[1] and Christian Lahusen[2]
Katholische Universität Eichstätt-Ingolstadt and *Otto-Friedrich-Universität Bamberg, Germany*

In this paper, we discuss sampling problems and strategies for process-generated data. To this end, we centre on the media discourse on unemployment as an example, and here particularly on news coverage by newspapers. This data provides us with a rich source of information, because newspapers are an actor, arena and archive of public discourses at the same time. Depending on how the target population is defined, different sampling problems arise. Process-generated data are biased during data production, because they are produced for other purposes and under different contexts than scientific research. Data preservation is biased, too, as data can decay or might be destroyed. These biases and the definition of target population reduce the choice of sampling procedures. Therefore, we suggest using sampling procedures usually applied when sampling qualitative data. We then present different strategies for applying these techniques for sampling newspaper data.

Key words: Multi-Staged Sampling, Adaptive Sampling, Target Population, Cases, Sampling Bias, Most Different Cases Design, Longitudinal Research, Time Units, Time Span, Validity, Information Accessibility

1. INTRODUCTION

Empirical research in the social sciences is dependent on good data. This also means that valuable research questions often cannot be answered due to the lack of data and / or deficiencies of available data sets. One possibility to circumvent these problems is using process-generated data as an alternative form of data retrieval and generation. In this paper, we will demonstrate typical sampling problems for this data type. To this end, we will focus on one specific type of process-generated data: reports on unemployment in German newspapers. We will first introduce the research question we use as an example. Then we will discuss why it is necessary to use process-generated data to address this question. We then examine three problems that have to be solved in order to sample process-generated data: Researchers first have to define the population. Second, they have to consider how data are biased during data production and selection. Using this information, they finally have to develop the actual sampling strategy.

2. EXAMPLE RESEARCH QUESTION: REPORTS ON UNEMPLOYMENT IN GERMAN NEWSPAPERS

Social scientists have developed a particular interest in public discourses on social problems (van Dijk, 1997). The development of public discourse on unemployment is one, yet, very suggestive example. In fact, since the 1970s, unemployment has been rising continually in most European countries. Thus, the last 30 years were characterised by heated discussions on how to

[1] Address: Ostenstraße 26, 85071 Eichstätt, Germany; e-mail: nina.baur@ku-eichstaett.de
[2] Address: Lichtenhaidestraße 11, 96052 Bamberg, Germany; e-mail: christian.lahusen@sowi.uni-bamberg.de

reduce unemployment. During this time, mainstream academic discourse has shifted from (Neo-)Keynesian to Neo-Liberal arguments. Today, three argumentative structures compete: Both (Neo-)Keynesians and Neo-Liberals want to reduce unemployment and strengthen capitalism at the same time. While Neo-Liberals want to reduce social security, Neo-Keynesians try to save the welfare state. A third group of theorists seek an alternative to capitalism (Baur, 2001).

While scholars have been analysing either *scientific* or *intellectual* discourses (Baur, 2001), it is as intriguing to know how *public* discourse has changed over time: How important is the topic "unemployment" in public discourse? Who participates in public contentions about this issue? Which arguments and solutions do participants favour? Does public discourse refer to values and norms? And are public debates changing over time? We have tried to answer these questions in two related projects. The first project is part of a comparative project financed by the European Union. The research team analysed German discourse on unemployment in the newspaper "Die Süddeutsche Zeitung" from 1995 to 2002 (for project design and codebook, see Giugni/Statham, 2002; for first results see Baum/Lahusen, 2004). In this paper, we will call this project "Project A". "Project B" was a local project conducted by Nina Baur and Christian Lahusen together with a team of students. We analysed news coverage in ten German newspapers between 1964 and 2000.

3. WHY PROCESS-GENERATED DATA IS NECESSARY FOR OUR RESEARCH QUESTION

Public discourses are a fairly new and interesting area of analysis. However, research on this topic is confronted by a serious problem: the lack of data. In fact, if – as in our example – the research question demands tracing change over several decades, researchers need data from these periods. Many other researchers face the same problem because neither survey data nor qualitative interviews are available or adequate for their particular research question. Some of the reasons for this lack of data are (Baur, 2004):

1) *New research questions may arise.* For example, until the 1970s, German employers actually had difficulties finding employees. Because of full employment, unemployment was no topic at all. In the 1980s, however, mass unemployment became a social fact and a public concern. Yet, only in recent years researchers became interested in public and scientific discourses on unemployment. As this topic had been regarded as mostly unimportant or uninteresting until then, no relevant data was collected. This will always happen: As society changes, social scientists' research questions will change.
2) *Over the past few decades, German sociologists centred on individual change*, e. g., on shifting individual attitudes to work, employment careers and/or personal fears of becoming unemployed. Nonetheless, discourses are collective phenomena. The units of analysis in the sense of sampling are not individual persons but claims about unemployment (for a definition of "claims", see Giugni/Statham, 2002). Put together, these claims form the discourse. In Germany, hardly any databases on discourses exist.
3) *For decades, German social scientists have focussed on surveys*. Questionnaires cannot answer certain types of questions. This is the case with discourses because they transcend the individual. Moreover, while issues and agendas might change more quickly, the argumentative or ideational basis of discourses evolves commonly over large time spans – years, sometimes decades. In this sense, we are dealing with short and long cycles of changes. Consequently, persons might simply not be aware of discourse details and/or developments on both levels.
4) *Social scientists have continually improved their methods over the last twenty years.* New types of questionnaires, new analysis procedures and new design types have been devel-

oped. For example, event history analysis and sequence analysis are fairly new procedures. If researchers wish to benefit from these methodological improvements, they might not be able to use older data because these data might be appropriate only for old-fashioned methods. For example, event history analysis demands event data. In Germany, event data has been widely collected since the end of the 1980s only.

5) *For cross-cultural comparisons, respective data sets need to be available for all countries.* For example, in Project A, different countries' discourses on unemployment were to be compared. Even if an appropriate dataset was available for one country, it is highly unlikely that a similar dataset exists for the other countries, especially as these datasets need to be comparable with regard to their content. Moreover, they also need to cover the same time span.

In our projects, lack of data could not be resolved by using retrospective interviews: We were interested in discourses as they evolve in the course of time. If we interviewed people today, several problems might arise: First, relevant participants in discourses might be dead or not traceable. Second, these participants might not remember all relevant details. Third, they would construct their version of the past from their present viewpoint. Relating to the discourse on unemployment, their opinion on things might have changed, they might not know the full story. Consequently, people will tell a different story today than they would have told twenty years ago.

In cases such as these, using process-generated data is an alternative. Process-generated data is data not produced for scientific research. Instead, it is the result or by-product of social processes. Examples are newspaper articles, contracts, laws, speeches, records, files, protocols, diaries, personal notes, emails, letters, websites, databases, internet protocols, clothes, commodities, tools, furniture, architecture, landscapes, photography, films, comics, paintings, sculptures, maps and so on. Thus, the array of documents that can be used as a source of information for scientific research is wide. They can be standardized, semi-standardized or not at all standardized. In our case, we used newspaper articles, which is a source of process-generated data that has rarely been employed so far (Müller, 1996). Usually, information from process-generated data can be transferred to a database using qualitative and/or quantitative content analysis. Alternatively, this information can be analysed using interpretative methods. Our research used a mixed-methods design because we both transferred the information to a standardized data base and analysed it using qualitative methods (for details see Lahusen/Baur, 2005).

Before analysing process-generated data, though, researchers have to collect them. Usually, the amount of available process-generated data is so high that they have to be sampled. In contrast to scholars relying on surveys and narrative interviews, researchers using newspapers (and other kinds of process-generated data) have rarely applied systematic sampling procedures so far (Lerg/Schmolke, 1995). Sampling strategies determine if and how results of data analysis can be generalised. In order to determine if a sample of data is a reliable source of information, researchers have to address three problems:

1) *Researchers have to define the target population and the cases this population consists of* (Behnke et al., 2004). As we will show in chapter 4, what is a fairly simple task for surveys, is much more complicated for process-generated data. The reason is the multiperspectivity of this kind of data.
2) *If researchers want to test hypothesis or calculate confidence intervals, they need random samples. Thus, the sample should not be biased* (Behnke et al., 2004). As we will show in chapter 5, process-generated data is often biased during the production and selection process. These biases depend on the data type and the target population.
3) *Using information about target population and biases, researchers have to develop an actual sampling strategy*, that is, they have to select and find relevant cases. We will show in

chapter 6 that a multi-staged sampling strategy is required. We will also show that different kinds of samples can be used to evaluate biases and thus make results more reliable.

4. TARGET POPULATION AND MULTIPERSPECTIVITY OF PROCESS-GENERATED DATA

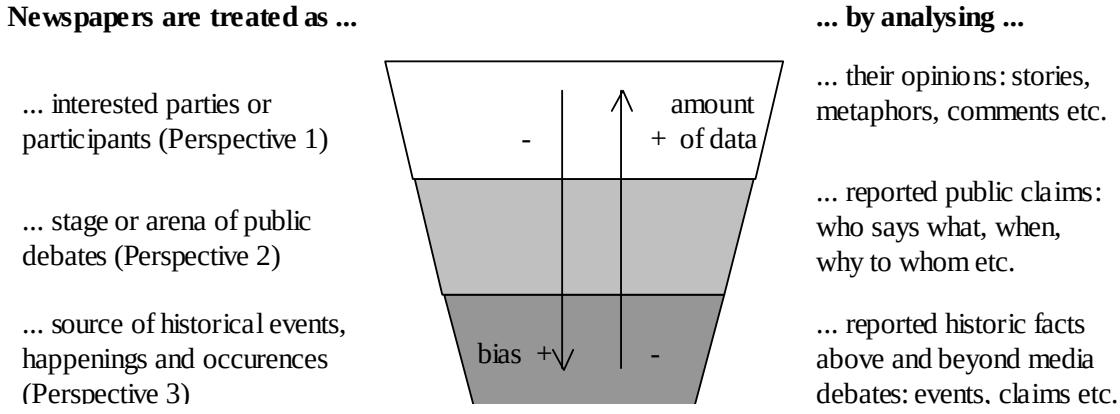
The first step in sampling is to define the target population (Behnke et al., 2004). For surveys, this is usually straightforward. Researchers, for instance, might be interested in analysing the attitudes of Germans in regard to unemployment in 2004. Depending on how you define the term "German", the target population consists either of all ethnic Germans (regardless of where they live) and/or of all persons living in Germany in that year. A case is one person belonging to this population.

In contrast, defining the target population for process-generated data is more complicated. The reason resides in the ambivalence of data – particularly mass media products can be read from different perspectives. For example, mass media evolved in modern society as an instrument to observe and report about reality. In Germany, in the 18th century historical science was divided into historical science and journalism. While the former was from then on responsible for tracing long-term development, the latter was responsible for writing "daily" history. Sharing the same roots, both occupations have been conceived as being bound to neutrality and seeking truth (Keppler, 2000; on the history of German newspaper system see also: Wilke, 1999). In terms of functionalism we can argue that the more societal actors and systems became dependent on mass media, the more did the value of newspapers consist in the production of reliable information. However, this is only part of the story (MacQuail, 2003). Mass media are also an actor or institution in their own right, with proper working routines and organizational needs. When we read newspapers or watch TV news, we do not get an (unbiased) picture of reality, but rather the journalists' highly selective view on social processes.

Both aspects are intertwined: Media are not a mirror of reality, although they are obviously a highly institutionalised form of describing reality. While theories of mass media have tended to highlight either the one or the other aspect, we argue that this ambivalence is a specific quality of news coverage. Texts produced by media are an amalgamation and conflation of different realities: they are a window to external reality and a institutionalized reality by themselves. Thus, researchers can read texts in multiple ways, depending on their specific perspective or research question. In this sense, we argue that newspapers can be read and analysed at least in three different ways (see figure 1):

According to *Perspective 1* we can analyse mass media in order to catch the story the journalist or the newspaper has to tell. In this case, *the journalist him-/herself or the newspaper itself* is our object of research: We treat the media as an actor who is heavily involved in public discourses and thus affects people's knowledge, opinions and attitudes. While newspaper articles are there to inform people about facts, events and claims of social actors, we would solely be interested in extracting the media's core message or story. Thus, we would focus primarily on the specific semantic organization of information, (implicit) comments and rhetoric devices (metaphors, catch words, examples etc) in order to capture what meaning and message the newspaper or journalist ascribes to reality. For instance, researchers would analyse how journalists and newspapers report on what other actors did and how they behave in the discourse (Pietilä, 1992).

Figure 1: Possible Perspectives when Reading Newspaper Articles



If researchers are interested in a journalist as a participant of public discourses, then the target population are all articles this journalist has ever written. If researchers are interested in a newspaper (for example "Die Süddeutsche Zeitung"), the target population are articles that have ever been published in the particular paper. If researchers are interested in more than one journalist or paper, the target population is what has ever been published by all respective journalists or papers. For all these examples, a case is one single article. If researchers are interested in Perspective 1, newspapers are a quite reliable source of information: As discussion in chapter 5 will prove, there is no data production bias, although there may still be a data selection bias. Indeed, the biased reflection of reality is not a methodological problem but rather the specific information we want to gather: here, the particular selectivity and point of view of a journalist or newspapers. At the same time, the amount of data is abundant. In fact, it is usually too abundant. As we will show in chapter 6, researchers usually can only choose a small sample of data for this case.

Following *Perspective 2* researchers conceive of mass media as a public arena or stage. Social scientists even argue that the mass media have become the most important turntable of the public space, of "Öffentlichkeit" in modern societies (McQuail, 1992). While face-to-face encounters and public meetings or events are a dominant form of establishing public spaces on the level of daily life, it is the media that provide a stage for public information and opinion formation on the macro-level of complex societies. If we opt for Perspective 2, we will be interested in analysing *public discourses*, as they unfold within the mass media. Although journalists may be part of these discourses (Kepplinger, 1994; Hoffmann/Sarcinelli, 1999), we will be interested in deciphering the (implicit) opinion and story of the journalist only 'negatively', i.e., in order to identify the bias and control for unwanted effects on our primary object of research: reported public discourses. For this purpose, we would centre our data retrieval on information about which actors have said or done what, when, why, and in view of whom. This enables us to reconstruct and understand the structure and dynamics of public discourses, as they form within the public arena of mass media.

In this case, the target population are all actors participating in a particular public discourse, for example, the discourse on unemployment. A single case is a claim made by one of these actors. As newspapers are an important arena of public discourse, a lot of information on public discourses can be found in newspapers. Again, only a small sample can be analysed. However, data is never complete: Only part of the overall discourse is reported on in newspapers. This is

due to the selectivity of news coverage (Bright et al., 1999). In fact, journalists are a major gatekeeper to public discourses (Shoemaker, 1991). Moreover, different newspapers report differently on the same discourse. Thus, researchers using Perspective 2 have to be aware of the data selection bias, but also partially of the production bias. Here, we touch a specific problem of the second perspective. In principle, the selectivity of mass media is not a methodological problem for this perspective *per se*, because we want to study mass mediated public discourses as one of the decisive arenas of public contention and debate. Hence, we are not interested in grasping the 'real', multifacetted and/or unorganized discourses out there, but the publicized, highly patterned and constructed debates within the media. These debates are biased, first, because mass media construct news along their working routines, and, second, because social actors consciously address the media in order to participate in this kind of debates and adapt their public appearances accordingly. Newspaper articles can be understood as a materialization of such double-faced structuration processes. In this case, the production bias is not a problem but an inherent part of the structure and dynamic of mass mediated public discourses, which we want to analyse. While this observation is true when referring to the mass media in general, methodological problems arise, however, when we decide to sample only a particular type of media (e.g., newspapers) and individual press organs (e.g., one daily newspaper). Each newspaper can be considered a public arena on its own, yet, mass mediated discourses extend to different 'arenas', e.g., different newspapers, TV channels, radio stations and internet pages. Hence, individual newspapers only provide a limited window to mass mediated discourses, which means that the chosen press organ is biased and thus not representative of the mass mediated discourse as such. Hence, we can conclude for Perspective 2 that an inquiry of all mass media is not biased, while this bias increases the more we reduce the number of media organs under scrutiny.

Finally, *Perspective 3* regards newspapers as a medium and/or as a (biased) window to *reality itself*. Press coverage is seen as an archive for historical facts (Franzosi, 1987), for instance, when we aim at analysing political debates within and around the political institutions (i.e., parliamentary discussions, negotiations between ministries, contentions along electoral campaigns, social dialogue between employers and unions and many others). In the case of political contentions about unemployment policies we would assume that the press gives us a more or less restraint testimony of these debates and conflicts – trusting that the media present us at least part of the story. The target population are all facts, events, occurrences and happenings regarding these conflicts. Single facts, events, occurrences and happenings are the cases. How much information on actual social processes (other than public discourses) researchers can extract from media data depends on the type of discourse: As we will show later on, one can draw quite reliable information on German official unemployment figures or strikes from German newspapers. It is a lot harder to find out how the unemployed themselves live and behave. All in all, *Perspective 3* is the perspective where the amount of data contained in newspaper articles is smallest, when compared with the many other sources of information we could use. Moreover, this newspaper data is most biased in regard to production and selection processes: Reality is broken twice – through public discourse's and through journalists' perspectives. As researchers are interested in facts, they can solve these problems only by triangulating data, e. g., by looking for reports from other newspapers, interviews, documents and so on (on triangulation see Seale, 1999; Flick, 2000).

In summary, newspaper articles reflect a multiperspectivity of reality: They can be read in different ways, drawing different kinds of information from them. This ambivalence does not apply uniformly to all media types. In practice, communications studies have centred on TV channels as actors of public discourses in order to unveil the impact of news coverage on public perceptions and opinions (e.g., Brooks, 2004; Chiricos/Padgett/Gertz, 2000; Tudor, 1992). TV is a better candidate for this kind of research because the selectivity and construction process is much more evident in the case of television than in the case of newspapers. It is no surprise that

newspapers are therefore more recurrently used when researchers analyse public discourses and/or are looking for archives of empirical data (Franzosi, 1987). In this paper we argue that newspapers mirror the above mentioned ambivalence or multiperspectivity more clearly than the other media. We thus concentrate on newspaper data because they can be analysed as actors, arenas and archives at the same time.

This multiperspectivity has implications for data sampling. Depending on the research question and perspective on the data, different kinds of target populations and cases have to be defined. Different perspectives in turn face different kinds of biases, have to handle different amounts of data and have to use different sampling strategies. This ambivalence does not necessarily create problems for empirical research, as long as we know what kind of information we can and want to ask from data, and as long as we develop strategies to control the inherent bias of this medium. Hence, from a methodological point of view, the latter aspect needs more consideration before sampling strategies can be developed.

5. RANDOM SAMPLES AND BIAS

The whole idea of sampling is to create a sample of data that allows to generalise findings after data analysis. One typical way of generalising is using hypothesis testing or confidence intervals. Both statistical procedures require random samples. In other words: If researchers want to apply inferential statistics, the sample should not be biased (Mayer, 1998; Gigerenzer, 1999; Behnke et al., 2004). For surveys, biases are usually determined by a bad research design, non-response and missing values (Schnell 1986, 1997). In panels, panel mortality and spell effects may additionally bias data (Blossfeld et al., 1986; Steinhage, 2000). Additionally, the target population may change over time (Abbott, 1995, Baur, 2004). Most of these problems can be handled or at least minimised by a careful research design. Therefore, surveys can, but do not necessarily have to be biased.

The situation is different when using process-generated data as an archive: Process-generated data is almost always biased. This means that the target population and the frame population differ. The frame population over-covers some types of cases of the target population, other types of cases are under-covered or even completely absent from the data (Behnke et al., 2004). Two intertwining processes might influence the bias: As researchers cannot control the process of data production, data is usually already biased during production. Moreover, this bias will accumulate over time: more and more original data might decay or will be destroyed deliberately. This process, too, is biased, as humans have to actively want to preserve data available for later use. The researcher can influence neither data production nor data selection process. However, both processes influence what data is available at all. Furthermore, these biases depend on the research perspective on newspaper data and on the target population: As long as researchers analyse newspapers as participants of public discourse (Perspective 1), data production bias is not a real problem but rather a peculiarity of this data type. As soon as we treat newspapers as an arena of public debates or as historical archives, data production bias becomes pertinent in relative (Perspective 2) and absolute terms (Perspective 3). In addition, researchers have to handle data selection biases for all three perspectives. Researchers cannot change this situation. However, learning from historians, social scientists might evaluate biases by qualitatively analysing production and selection contexts. Before demonstrating these possibilities by using our data as an example, we will discuss the reasons for data production and data selection biases in more details.

5.1 Biased Data Production Process

Data is supposed to document events and processes. Because so much is happening at the same time, a full account of reality is impossible. This is true for both process-generated data and for survey data. However, in contrast to surveys, researchers cannot control the process of data production when using process-generated data because the latter has been generated for other, practical purposes, e.g., to inform people on a daily basis on local, national or international news. For this reason, we have been arguing that process-generated data is usually biased during data production already. But *how* is it biased? Researchers need to answer this question in order to be able to develop adequate sample strategies. For the case of mass media, we argue that there are three elements that have a strong impact on the production bias:

1) *Purposes:* News media have publicly acclaimed rights and duties. In general, German journalists pursue the professional goal of informing the public, contributing to the formation of public opinions and controlling state power, e.g., by means of "investigative journalism". This professional self-concept applies particularly to the press, which stresses its public mandate quite clearly (MacDevitt, 2003; Keppler, 2000). Rules and checks have been instituted both on the level of daily working routines (e.g., the separation of information and commentary in news coverage) and on the level of institutional controls (e.g., liabilities of newspaper publishers) in order to ensure that these purposes are met effectively (Pöttker, 2002). At the same time, we have been arguing that the media accomplish this task in a highly patterned and selective way. This production bias tends to increase over time. In fact, German mass media are increasingly dominated by economic rationality: Globalisation and concentration increase competition within global and German mass media (Altmeppen, 2000, Küng, 2001; Wittenzellner, 2000),. Due to economic pressures, professional ideals are forsaken very often. The boundaries between information and entertainment (Wolf, 1999) as well as the boundaries between journalism and public relations are increasingly blurring (Trappel et al., 2002). The few studies existing on this topic suggest that journalists all over the world are far from the ideal of investigative journalism because they do not know enough about the topic, they do not work continually enough on the same topic in order to build up knowledge, and because they do not have the time for their own research (Bow, 1980). In this sense, we are dealing with an increasing variety of purposes that affect the production bias. That is, when dealing with mass media as public arenas or historical archives we have to consider the particular orientations and purposes of the chosen media.
2) *Formats:* Newspapers conform to various formats. On the level of the individual article we know that information is selected according to specific rules (e.g., news values, see Galtung & Ruge, 1973) and that articles follow a particular pattern and structure (e.g., summary, main event, background, comments; see van Dijk, 1988). On the level of individual newspaper issues we are dealing with different sections (e.g., national and international news, business, culture, sports), which have quite different agendas and a particular approach towards news coverage. This means, for instance, that unemployment will not rank high in sport or culture sections. Additionally, reports on unemployment will have an entirely different story to tell in these sections, when compared to the pages covering political or business news. On the level of newspapers we may distinguish between prestige newspapers and tabloids, daily and weekly, local and national newspapers, papers with a broad scope of topics and highly specialized newspapers, amongst others (Kepplinger, 1994; Wilke, 1999; see also table 1). Each of these newspapers has a specific agenda and orientation (Oliver/Maney, 2000; van den Berg et al., 1992) that might change over time, yet, unveils a certain identity and continuity. On all these levels, media formats are highly patterned and institutionalized. This means that we can get acquainted with them. In this case, we will know which article, sec-

tion or newspaper we have to read when we are interested, for instance, in information, opinion or entertainment on local or national news, politics or culture and so on.

3) *Institutional embeddedness*: On a more general level, the production-based bias is determined also by the environment of mass media. First, newspapers have to conform to legal guidelines of what and how to report (e.g., restrictions on pornography, respect of privacy). Second, most media (TV channels, radio stations, but primarily newspapers) have quite stable political allegiances (see table 1). Moreover, the fact that newspapers depend on valuable sources (in most of the cases within government, parties and interest groups) brings about political loyalties or solidarities. Finally, mass media also depend on markets in that they are interested in securing or expanding the range of advertisers and readers (for the German case see Schulz, 1999). All these institutional environments have an impact on news reporting, e.g., by privileging certain topics, positions, formats or purposes to the detriment of others. In this sense, institutional factors are intertwined with journalists' purposes and formats: Due to increasing competitive pressure, journalists today have less time for a single article. They increasingly use companies' and political parties' press releases instead of their own research for articles. Thus newspaper articles tend to be biased towards specific actors. Also, institutions have their own rhythms: Election cycles, typical dates for press releases, parliamentary sessions and so on. This means, there are more news on some topics at specific dates, less on others. This affects reports on other topics, as the amount of slots in a given newspaper is limited (Oliver/Maney, 2000). "Unemployment" is a high priority topic in Germany, thus "stealing" slots from other news. However, there are exceptions: Catastrophes, wars, elections and other news with a high news value crowd out other news (Oliver/Myers, 1998). They might also push "unemployment" to second place.

Table 1: Some German Newspapers

Newspaper	Rhythm of Publishing	Readers	Political Alignment
Der Spiegel	Weekly	National	Liberal Left
Die Zeit	Weekly	National	Liberal Left
Die Frankfurter Rundschau (FR)	Daily	National	Liberal Left
Die Süddeutsche Zeitung (SZ)	Daily	National	Liberal Left
Die Bild-Zeitung	Daily	National	Conservative
Die Frankfurter Allgemeine Zeitung (FAZ)	Daily	National	Conservative
Die Welt	Daily	National	Conservative
Fränkischer Tag (FT)	Daily	Regional	Conservative
Nürnberger Nachrichten	Daily	Regional	Social Democratic
Stuttgarter Zeitung	Daily	Regional	Conservative

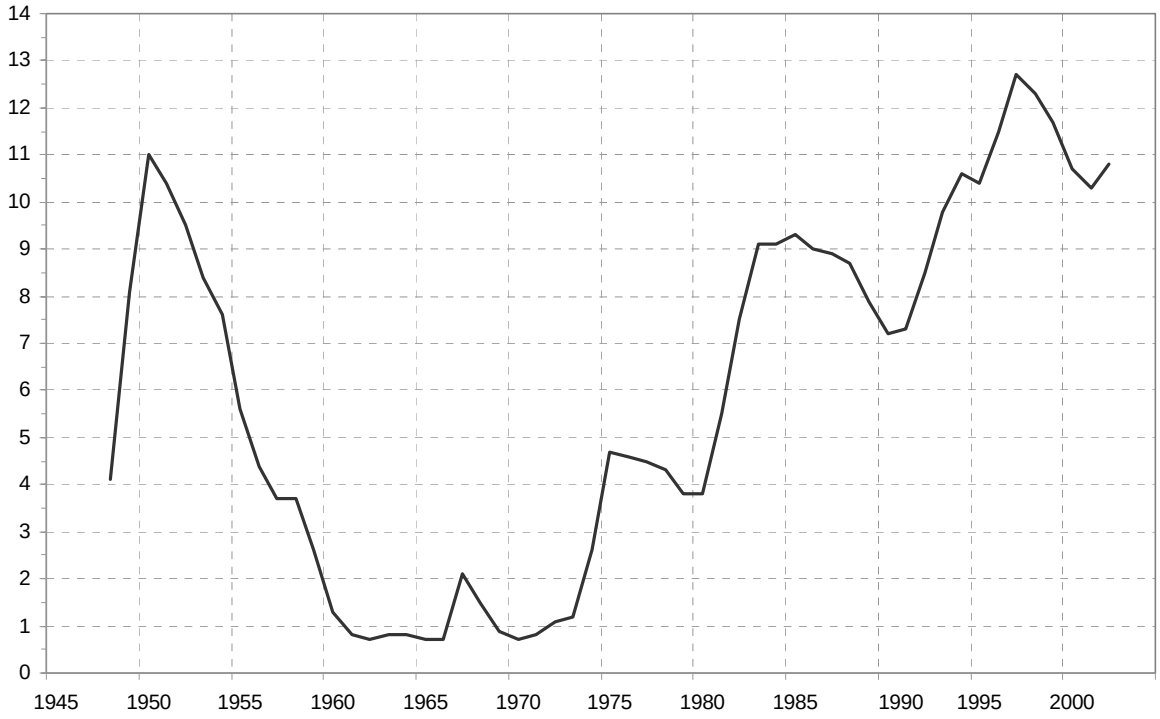
In summary, newspapers cover issues differently depending, e.g., on time of occurrence, location of occurrence and the relevance of the issue for the social context (Hocke, 1998; McCarthy et al., 1996; Oliver/Maney, 2000). In order to assess selectivity, researchers therefore have to analyse the social context. For example, we consider the following types of context information as especially important for the amount of news coverage (and therefore for production bias) on the German discourse on unemployment:

1) *The Development of Unemployment:* Discourses on unemployment are not independent from the actual development of unemployment (see also figure 2). In Germany, the 1950s and 1960s had been earmarked by full employment. Actually, there was a lack of labour. Since the middle of the 1970s – starting with the oil shock – unemployment has been rising continually. After unification in 1990, unemployment figures jumped (mainly due to high

unemployment figures in East Germany). This means, that in the 1950s and 1960s public discourse centred on the lack of labour. One can extract almost no information on unemployment from German newspapers for this time span. In contrast, unemployment has been a central topic in German newspapers ever since the 1980s.

2) *Historical Legacies:* In Germany, unemployment is a sensitive topic: During the 1920s and 1930s, the social climate was very instable. Then unemployment rose rapidly during the recession in the 1930s. This combination made it possible for Hitler to win the elections. Thus unemployment is usually conceived of as an immediate cause of destabilising democracy. This was especially important in 1997, when unemployment figures topped the unemployment figures of the 1930s for the first time since 1949 (Baum/Lahusen, 2004). Additionally, unemployment always receives high coverage in comparison to other topics.
3) *Political System:* The political system also influences the way public discourses work. For example, it is likely for election years and election campaigns to influence such debates, both in their intensity and internal structure. As unemployment is considered such an important topic, elections mean that there is even more news on unemployment, while other topics are crowded out. However, the character of the discourse changes during this time: Politicians are even more dominant in the discourse than usually (Schmitt-Beck/Pfetsch, 1994). During election times, discourse is centred more on staged events ("Pseudoereignisse") and on seeking culprits (for high unemployment) than on finding manageable solutions (Schmitt-Beck/Pfetsch, 1994). Public discourses do not only reflect election cycles but also

Figure 2: Official Unemployment Rate for Germany (in %)



Definition of unemployment and/or measurement methods were changed in 1966, 1985, 1987, 1989 and 2004. The figures also measure unemployment rates for different regions: Until 1949 for West-Germany without Saarland and West Berlin; from 1950 to 1958 West-Germany without Saarland; from 1959 to 1990 for West Germany; since 1991 for West and East Germany

Source: Bundesagentur für Arbeit, 24.05.2004, http://www1.arbeitsamt.de/hst/services/statistik/aktuell/iiia4/zr alo qu ab 1948d.xls

the federal structure of German polity. This means, for instance, that regional papers (or regional sections in national papers) become an important source of information when analysing regional debates. German news coverage is also known to be government centred. We thus need to know the succession of parties in government in order to estimate the effect of this bias on the inclusion or exclusion of political parties from news coverage (see table 2). Additionally, it is important to know that labour market policies are institutionalised in a neo-corporatist way: Trade union associations and employers' association decide on many areas of labour market policies. Different newspapers and different newspaper sections will provide us with quite different information on these organizations. For instance, business sections and business newspapers usually favour employers' organisations (Gesterkamp, 1993). Finally, German welfare associations are not represented in labour market policies, while being heavily involved in social policies and, thus, in the practical work with the unemployed. If we want to reflect the public discourses and contentions in this wider spectrum, we would need to select articles on unemployment and labour market issues and on social security.

4) *Social Security Systems:* Actors will not talk about things they take for granted, only about issues debated. What is taken for granted and what is debated depends on the specifics of the social security system, as every welfare state targets certain social problems but is also a source of new conflicts. In other words: Just because people do not talk about certain problems, this does not mean they do not exist. On the other hand, certain problems might be framed in a different context. Since the 1920s, Germany has had an extensive system of unemployment benefits. The national unemployment agency (formerly "Bundesanstalt für Arbeit", now "Bundesagentur für Arbeit" (= "BA")) also collects statistics on the development of employment. Germans thus conceive of a person as "unemployed" if they are reported to be unemployed at the BA, i.e., as a person looking for employment. Thus, labour market dif-

Table 2: German Governments since 1949

Chancellor	Governing Period	Governing Parties	Elections
Konrad Adenauer	September 1949 – October 1963	Conservatives (CDU / CSU) & Liberals (FDP)	14.08.1949
			06.09.1953
			15.09.1957
Ludwig Erhardt	October 1963 – December 1966	Conservatives (CDU / CSU) & Liberals (FDP)*	17.09.1961
			19.09.1965
Kurt Georg Kiesinger	December 1966 – October 1969	Conservatives (CDU / CSU) & Social Democrats (SPD)	28.09.1969
Will Brandt	October 1969 – May 1974	Social Democrats (SPD) & Liberals (FDP)	19.11.1972
Helmut Schmidt	May 1974 – October 1982	Social Democrats (SPD) & Liberals (FDP)	03.10.1976
			05.10.1980
Helmut Kohl	October 1982 – October 1998	Conservatives (CDU / CSU) & Liberals (FDP)	06.03.1983
			25.01.1987
			02.12.1990
			16.10.1994
Gerhard Schröder	October 1998 – ?	Social Democrats (SPD) & Green Party (Bündnis 90 / Die Grünen)	27.09.1998
			22.09.2002

* Not all the time.

ficulties of a lot of population groups are not debated as problems of unemployment. For example, Germany is a Conservative Welfare Regime (Esping-Andersen, 1990) with a Strong Male Breadwinner Model (Ostner, 1995). If women are laid off, very often they drop out of the labour market completely. Unemployment is considered mainly a male problem – by both men *and* women. Young people either go to university or they participate in the well-developed system of occupational training ("Ausbildungssystem"). Actors therefore never talk about youth unemployment but about a lack of trainee slots ("Ausbildungsplätze"). Unemployment security only covers certain occupational groups: people employed by companies ("abhängig Beschäftigte"). State employees ("Beamte") have a guaranteed life-time employment. The self-employed ("Selbständige") are not part of the social security system. Therefore, neither group is talked about in the context of unemployment. The same is true for ethnic minorities, though for different reasons: There is no separate employment statistic on ethnic minorities. Germany's citizenship is based on *ius sanguinis*. Therefore ethnic Germans – regardless of whether they were born in Germany or are immigrated – are considered Germans (German Information Center, 1995). Persons seeking asylum are not allowed to work. If they are discussed at all, it is in the context of moonlighting. In the 1960s, Germany invited immigrants from Southern Europe (especially Turkey, Spain and Italy) to work in Germany as "guest workers" ("Gastarbeiter"). The idea was to send them back once they were not needed anylonger. Of course this did not work. Today, the guest workers are either integrated in the labour market (having paid into the social security systems and thus being eligible to benefits) or have returned to their country of origin (Castles/Kosack, 1985). All this information is important for deciding how the newspaper data are biased. For instance, if researchers want to analyse discourses on youth unemployment in Germany, they would have to consider articles dealing with the training system, while this might not be necessary in other countries.

5) *Time Span of Important Discourse Changes:* We have argued that public discourses evolve according to short-term and long-term cycles: while news coverage changes quickly, when issues and agendas are concerned, this is not true when looking at underlying ideas, ideologies or world-views. For instance, academic discourse on unemployment shifted at the beginning of the 1970s from discussions about the distribution of labour to debates about the preventive avoidance of unemployment. Until the mid-1980s, academic discourse was dominated by a labour-friendly, Keynesian position. Since then it has been shifting to an employer-friendly, neo-liberal position (Baur, 2001). It is likely that media discourse shifted correspondingly. In case we want to grasp these shifts, we would need to generate a sample of data that covers this longer time span.

These observations illustrate that mass media data is biased, if researchers use Perspective 3, and to a lesser degree Perspective 2. This means that it is impossible to draw random samples for research questions using these Perspectives, although one might draw random samples for Perspective 1, and to a certain extent also for Perspective 2, in case we consider all mass media. However, this problem does not discredit process-generated data as a source of information. Instead, researchers can use alternative sampling strategies that are typically used in qualitative research (for an overview see Creswell, 1998). Examples are the selection of typical cases, the most different cases design and the most similar cases design (Behnke et al., 2004). We will demonstrate this in more detail in chapter 6.

5.2 Biased Data Selection Process

So far, we have discussed how data production biases affect process generated data. When using this data in order to trace social change, researchers face an additional problem: Data has been

produced at earlier points in time than it is analysed. During the time that has passed between data production, data preparation and data analysis, data can be lost. What appears harmless at first sight may seriously impair research results as data selection is biased, too.

First of all, human beings might deliberately destroy data because they do not want the information being kept for former generations. This happens to documents very often. A recent example is the German governmental change in 1998. When the Social Democrats took over the "Bundeskanzleramt" (the Chancellor's personal administration), they discovered that obviously important files were missing. Until today we know that there is something missing. However, we do not know exactly what files are missing and what information they contained. It is very unlikely that random data was destroyed. Probably, the missing data contained some information unfavourable to the former government.

Another reason why data might be destroyed is to make room for other data: Newspapers are published daily. They soon pile up and occupy a lot of storage room. Individual readers and researchers might decide therefore to throw away their personal archives. Even public libraries, archives or publishers might choose to do so if data has not been used a lot or is considered less valuable at that period of time. Moreover, data might get lost as a consequence of the bankruptcy of publishers or because archives are closed down and/or assigned to other entities. Finally, natural disasters or social turmoils might have their share in the destruction of data. In the case of newspapers, data destruction is generally no problem: Newspapers usually do not contain secret but public information. In addition, so many copies are printed and distributed that it is very unlikely that *all* of them were destroyed.

A second factor leading to a data selection bias is the decay of original data. For example, newspapers are usually printed on paper. Paper may burn. It may get wet and rot. The ink may dissolve or eat the paper, and so on. Therefore, people have to take active measures in order to prevent data from deterioration. They only take such measures for things they consider important. Here again, we could assume that the danger of decay might not be a severe one, bearing in mind that newspapers are mass products present in a number of different archives. However, while we have no empirical evidence, it is to be assumed that the danger of data destruction and decay affects less prestigious and more short-lived newspapers. We would expect that highly recognized newspaper will hardly disappear from human memory, while this might not be the case with ephemeral papers or tabloids.

However, newspapers are a commodity for everyday use. Therefore, not *all* but *most* copies of newspapers probably have been destroyed. Thus, the difficulty with newspapers is finding the remaining copies. Researchers can rely on a variety of different sources. In most European countries, for instance, newspapers have been archived for at least 100 years either by the publishing companies themselves or by public agencies. In Germany, each federal state has a central library or archive (the "Staatsbibliotheken" and "Staatsarchive"). These archives collect one copy of everything that has been printed in the respective region. Important German newspapers are also archived in university libraries and public libraries. Some archives and libraries collect newspapers or articles on certain topics as well. For example, staff of the "Hamburger Weltwirtschaftsarchiv" (HWWA) have been reading several hundred German newspapers every day since the 1970s. From these papers, they have been selecting all articles on companies or industries. Thus, for each company a file exists containing (almost) all articles that have been written on this company in Germany since 1970. In recent years, major newspaper producers have started digitalising their newspapers. Search machines can be used to find articles quickly. The articles can then be easily printed or imported into CAQDAS. This sampling method was used for Project A. However, only a minority of newspapers publish their articles on CD-ROM. In addition, so far data has only been available since the beginning or middle of the 1990s. Thus, for Project B, we had to draw on libraries or microfilms in order to obtain the relevant information.

Although newspapers can be stored in multiple formats and locations, this does not have to be the case. Prestigious papers are present at various locations and archives. In fact, many public libraries and archives only collect major and some local newspapers. Moreover, these papers are available in various forms (paper issues, microfilms, CDs), while tabloids are less accessible in terms of the number of archives and well developed electronic data-bases. For example, we had to rely on a mixed sampling strategy for Project B: The major newspapers ("Die Zeit", "Der Spiegel", "Die Süddeutsche", "FAZ", "FR", "Die Welt") were available either in the Bamberg University Library or in the "Staatsarchiv" which is also situated in Bamberg. For the local newspapers, we actually had to drive to the company archives, which is one of the reasons we selected *these* newspapers: "Fränkischer Tag" is situated in Bamberg, "Nürnberger" in Nuremberg and "Stuttgarter Zeitung" in Stuttgart. These newspapers were thus easily accessible. We had problems in obtaining relevant articles for "Die Bild-Zeitung": Being a tabloid and being printed in Hamburg, "Bild" is not archived in any place close to Bamberg and no electronic data base is available. Accessibility is certainly not a severe problem, however, it does constitute an implicit selection bias. Researchers have to know the social context well in order to know *where* process-generated data is stored, how to find it and how to gain access to it.

6. STRATEGIC CHOICES FOR SAMPLING STRATEGIES

After having defined the target population and after assessing data production and selection biases, researchers have to develop the actual sampling strategy. This third step is necessary because mass media provide data in abundance. Researcher interested in the analysis of process-generated data must develop an adequate strategy for reducing the amount of data without loosing the possibility of generalizing findings. So far, we have been arguing that these three steps cannot be separated because they are inherently intertwined: The choice of a sampling strategy depends on the specific research question, while the latter has also consequences for the kind of information we extract from newspaper articles. As we have shown in regard to the issue of multiperspectivity, the particular approach towards newspaper data determines the definition of target population and cases, and impinges on whether production biases are treated as a methodological problem of empirical finding.

When using newspaper data for Perspective 1, production biases are generally not a problem. Therefore, researchers can draw multi-staged random samples for these perspectives. The procedure is the same as for surveys (e. g. Cochran, 1972). The only difference is that articles or claims are sampled instead of persons. Of course, researchers can also use the alternative sampling strategies we will demonstrate below. In fact, if only a small number of cases can be sampled, this is advisable. When using newspaper data for Perspective 2, we have to consider production biases, the more we restrict our sample to a particular group of media organs. In regard to Perspective 3, data is always biased. In this case, random samples do not make sense: Biases either over-layer random error or completely render inferential statistics impossible (Mayer, 1998; Gigerenzer, 1999; Behnke et al., 2004). In these cases, we suggest using sampling strategies typical of sampling qualitative and historical data (for an overview see Creswell, 1998; Behnke et al., 2004).

What has to be taken into account for sampling? First, the unit of analysis is either a single article (Perspective 1) or an event or claim stated in this article (Perspectives 2 and 3). In order to find the units of analysis, researchers first have to select relevant newspapers and relevant issues. Thus, generally a multi-staged sampling strategy is required for sampling process-generated context: Researchers first have to decide which newspapers to analyse (Stage 1). They then have to find which issues are to be analysed from these newspapers (Stage 2). From each

issue, relevant articles have to be chosen (Stage 3). For Perspectives 2 and 3 events, facts and claims have to be identified within each of these articles (Stage 4).

Second, in addition to defining the target population and assessing bias, researchers have to decide how they want to locate their sample on the following dimensions:

1) *the number of newspapers to be sampled*;
2) *the time span, i.e. the number of years to be sampled*;
3) *the number of issues to be sampled per newspaper and year*;
4) *the number of articles to be sampled per issue*;
5) (only for Perspectives 2 and 3) *the number of claims or events per article*.

For each of these dimensions, the strata stretches from single-case design (one newspaper; one year; one issue; one article; one claim or event) to sampling the whole population (all newspapers; all years of interest to the research question; all issues that have been published by all newspapers in the selected years; all articles in these issues; all claims or events cited in all these articles). The result is a five-dimensional space of sampling possibilities. The extremes in this space would be a single claim, event or article versus all claims, events or articles that have ever been written (about) in Germany.

While being aware of these extremes is helpful for making strategic choices on sampling, neither extreme is suitable as an actual sampling strategy. One the one hand, one cannot generalise from a single claim, event or article. On the other hand, researchers can only spend limited time and money on answering a single research question. Usually, researchers have to define the maximum number of articles, claims or events that can be analysed in a given period of time. This maximum number does not depend merely on funding but also on the planned analysis strategy: It is lower for qualitative methods than for quantitative methods. Given this maximum number of articles to be sampled, researchers face a trade-off: Going up on one dimension means going down on another.

Researchers could still try to cut down dimensions. However, we argue that this has its price: data samples might become worthless or biased for particular kinds of research questions. Hence, the only possibility is to reduce dimensions that are not needed for answering one's own question. Or put differently: for particular types of research questions high scores on specific dimensions are important. For instance, a long time span opens a window to longitudinal analysis; the higher number of newspapers allows to consider a wider spectrum of arenas, orientations and positions; the bigger number of issues per year enables us to reconstruct discursive episodes more adequately; and the rising number of articles makes it possible for us to analyse debates at the intersection of various issue fields. In the following, we want to demonstrate how researchers can decide on how to reduce dimensionality. For illustrative purposes we will refer to our two research projects mentioned above. Note that there is no general solution to this problem. It can only be solved for specific research questions. Researchers need a definition of the target population. They also need to know the social context and how bias works.

6.1 Number of Newspapers

The number of newspapers has to be high, if researchers are interested in the position of different newspapers (Perspective 1) or in a wide spectrum of arenas (Perspective 2). If historical facts are to be reconstructed (Perspective 3), it is also important to sample as many newspapers as possible, as this allows researchers to triangulate different newspaper reports. If one of these aspects is important, it is sensible to make use of a most different cases design.

For example, when analysing discourses on unemployment, it is convenient to sample newspapers that differ as much as possible on the following parameters: (a) location of production; (b) regional vs. national scope; (c) publication cycle; and (d) political orientation. We used

this sampling strategy in Project B, selecting from the papers in table 1. Note that we decided to concentrate on general newspapers, dropping TV, the radio, the internet, books and special interest newspapers and journals from our frame of analysis.

In contrast, the team for Project A was not interested in variety of discourses but rather in a "typical" discourse. Thus, the research team chose only one paper: "Die Süddeutsche Zeitung" (SZ). The decision to use one newspaper only reduces the spectrum of news coverage and thus increases the production bias. The SZ is one of the national prestige newspapers, and therefore particularly interested in covering national political news. It is considered a moderately liberal newspaper, and thus represents the political mainstream with a weak leftist turn (see also Zakrzewski, 1995). This orientation provides us with a bias that fits well the research priorities of the project: The SZ is strongly focussed on the core policy domain (government and opposition, social partners, experts), while having a certain affinity to leftist organizations (e.g., the unions, welfare organizations, unemployed initiatives). As our project was interested in describing the fate of the unemployed and their organizations within the public discourse on unemployment, this choice seemed to be perfectly justified. Finally, the SZ has strong regional roots, as have all German national newspapers. By choosing the SZ, we opened a window to the federal structure of German polity, in this case to Bavaria, which plays a crucial role in German politics. Bavaria plays the role of an unofficial, Christian Democratic counter-government to the Social Democrats, who have been in power in Berlin since 1998. In spite of these regional roots, we established that news coverage is less regionally biased than within other national newspapers thus providing a balanced relationship between national and regional orientations, as the SZ is generally considered one of the most neutral national papers. This allowed increasing scores on the other dimensions for project A (see below).

6.2 Time-Span and Number of Years within this Time-Span

The choice of a time-span depends on the question of whether researchers are interested in cross-sectoral or longitudinal analyses. In the case of discourse analysis, researchers also have to decide whether they are more interested in short-term or long-term discourse cycles (Perspectives 1 and 2). While news coverage changes quickly, when issues and agendas are concerned, this does not apply when looking at underlying ideas, ideologies or world-views. When analysing issue cycles, we would decide to investigate a shorter period of time in more depth, while doing the opposite when studying the development of ideas or ideologies. In both cases, researchers need to decide on time spans on the basis of assumptions about issue and ideological cycles.

If researchers want to reconstruct historical facts (Perspective 3), however, a different sampling problem emerges: both the number of newspapers and the number of issues per newspaper and year has to be necessarily high. Thus, analysis is very time-consuming. For practical considerations this implies that the time-span to be analysed will generally be rather short, not least of all to allow triangulation between different data sources. There might be some relief, if researchers can use existing archives that collect articles on specific papers (e. g. the HWWA for German business news). However, this solves the problem only partially.

Our two research projects followed different choices in regard to the time-span. Using the context information discussed in chapter 5, we assumed that discourse on unemployment shifted at the beginning of the 1970s from discussions about the distribution of labour to debates about the preventive avoidance of unemployment. Till the mid-1980s, discourse seemed to be dominated by a labour-friendly, Keynesian position. Since then it has been shifting to an employer-friendly, neo-liberal position. In Project B, we wanted to grasp these ideational shifts. Thus we decided to generate a sample covering the period between 1964 and 2000. However,

limited project time made it necessary to sample only ten years out of the whole time period: Starting from 1964, we selected only issues from every 4th year in order to eliminate effects of election and economic cycles (see table 2). We tried to choose a year in the middle of legislative periods in order to avoid effects of election campaigns. This 4-year rhythm is slightly disturbed: The 1972 and 1983 elections were predated due to political crises. We also chose these particular years to make media data more or less comparable with survey data: For the years 1984, 1991 and 2000, some ALLBUS questions measure attitudes to personal economic success, to people in need (including the unemployed) and to social security (including unemployment security). The same is true for SOEP for the years 1987, 1992, 1997 and 2002. Due to the four-year cycle, we do not have data for all, but at least for some of these years. This rules out the possibility of identifying short-term shifts in discourse. Similar to panel data, spell effects may occur.

In contrast, Project A was interested in short-term issue cyles and complementary medium-term discourse shifts. Thus, a shorter time span was investigated (1995 to 2002). In contrast to Project B, *all* years were chosen thus ruling out bias concerning this dimension.

6.3 Number of Issues per Newspaper and Year

Scholars are sometimes interested in grasping the flow of events or claims, debates or conflicts more closely, e.g., in regard to discourses, negotiations, conflicts and so on. In this case, they have to increase the number of issues per newspaper and year, ideally by consulting the chosen newspapers every day. This is important because what we perceive as political contentions or debates at a given period of time, is made up of a succession of actions or speech acts, occurrences and reactions, which are reported within the media as 'news' on a daily basis. Thus, a large number of issues per year has the advantage of enabling researchers to reconstruct discursive episodes: Researchers can trace who reacted how to what; they can identify discourse communities, how arguments are twisted in discourses, how news holes are patterned and so on. If researchers want to use event history analysis or sequence analysis of short-term discourse, it is necessary to sample *all* issues from a given time-span. This observation is true for all three perspectives, although each of them focuses on a different aspect of these discourses. However, the number of issues researchers can analyse per newspaper and year correlates negatively with the number of newspapers and years selected: The more issues researchers want to analyse for a given period, the fewer newspapers and shorter time-spans they can analyse pragmatically.

Both our example research projects limit possibilities to trace discourse patterns and short-term changes: As eight years were analysed in Project A, the research team had to restrict the number of issues to three papers a week (Monday, Wednesday and Friday of each week). This makes it more difficult to reconstruct the thread of events and claims and thus the interactivity of public discourses. That is, it is not possible to analyse how individual actors interact on a daily basis. Still, the general patterns and dynamics do emerge from the data quite clearly. Moreover the sample allows to trace changes of discourse topics, opinions on a weekly basis, prominence of certain actors and so on.

In contrast, the Project B sample only allows to trace long-term discourse changes and broad tendencies because we could only sample year-wise. The first reason was that data collection was more time-consuming because we actually had to collect articles for ten newspapers from different archives, while Project A data could be collected from the CD for one paper only. Second, time and resources were far more limited for Project B compared to Project A.

The decision to select only a very reduced number of issues generated the problem of deciding which days to choose. A preliminary result of Project A was that coverage of unemployment issues was more or less the same all over the year (at least since 1997), with some exceptions:

First, summer is a weak period for news coverage in general as most Germans in general – and German politicians in particular – are on vacation. Second, the government is usually evaluated in public after its first 100 days in power. This provides an opportunity for debates on unemployment. Finally, German discourse on unemployment is highly ritualised in the sense that unemployment figures are publicized by the central unemployment agency (BA) in regular press conference. Regularly, this gives discourse on unemployment a new impetus for a couple of days, raising the number of reports during this time. Thus we selected the day after the press conference as reference day. Starting from the reference day, we read each issue of the respective newspapers until we found at least one article on unemployment. As unemployment is highest in winter and as German business analysts usually evaluate the first quarter of the year and forecast business development for the rest of the year in March, we chose the press conference at the end of March or beginning of April.

The next step was to find out the exact dates. We know that the press conferences have been held since the 1960s. With the help of Ms. Heidelies Künzel, a BA employee, we were able to trace back the dates to 1974 (see table 3). Starting from this information, we tried to reconstruct the earlier dates: The BA has the March data at its disposal on the first Monday or Tuesday in April. The press conferences are usually held the following day.

First analysis of these single issues showed that they contained some though not enough information about public discourse, particularly for the earlier years. The data did not reveal whether this was due to the fact that unemployment did not constitute a contentious issue at that time, that media coverage style was more concise and officious, or whether those press conferences did not provide a strong stimulus for public debates. For these reasons, we tested different adaptive sampling strategies in order to extent the range of articles sampled.

On the one hand, we used the "Deutscher Zeitungsindex", a printed data base of news coverage by national prestige papers. This index includes bibliographic references to prominent articles about the most various issues (amongst them unemployment and labour market) for a number of years. This provided us with an easy access to substantial reports. However, the "Zeitungsindex" is available only from 1974 until 1990, an important, yet incomplete time span. Moreover, no regional newspapers and tabloids are included. Finally, selection is highly reduced and selective.

On the other hand, we had to take up the more tedious work of consulting CD-Rom data bases from 1995 onwards and going through the paper versions for the remaining years. The goal was to assemble all articles on unemployment from two latter issues. Due to this time consuming work, we have not yet finished the data gathering process for all newspapers.

The above discussion shows that deciding on the number of newspapers, years and issues to be analysed, affects both sampling Stages 1 (selecting newspapers to be analysed) and 2 (selecting and finding relevant issues). Although these three dimensions and two stages can be separated theoretically, they are inseparably intertwined in actual research process. Together, they form the first sampling phase. The second sampling phase consists of decisions on the number of articles to choose from a single issue (Stage 3) and (for Perspectives 2 and 3) the number of claims or events to choose from a single article (Stage 4). Stages 3 and 4 are intertwined as well.

6.4 Number of Articles per Issue

The number of articles per issue is an important topic when it comes to deciding on the specific focus of one's own research object. On the one hand, we might be interested primarily in the discourse about unemployment in a strict sense. Here, it would be enough to select those articles addressing unemployment explicitly as the primary theme (e.g., header or lead). On the other hand, we could be focussing this public discourse in a comprehensive way. In this sense, we had

argued before that unemployment in Germany is not always debated under this heading, but also in regard to neighbouring issue fields (see chapter 5) such as social security, collective bargaining and companies' competitive behaviour, or under different labels as, e. g. occupational training (youth unemployment) and immigration (ethnic differentiation of unemployment), early pension schemes and retirement (elder and long-term unemployed). Moreover, it might be necessary to link different discourses with each other in order to better understand the meaning and structure of the unemployment debates, e.g., also by tracing back the diffusion of arguments, catchword or ideas.

When deciding on how many and what articles to choose from selected newspaper issues, newspaper can either draw random samples of all articles, or they can read all or selected newspaper sections and sample all articles relevant to the research question. For both Projects A and B, we opted for the latter procedure. All decisions were documented in a codebook (Giugni/Statham, 2002). We based these decisions on what we know about the German newspaper system and the context of German discourse on unemployment (see chapter 5).

First, we defined which sections were to be analysed. Drawing on our knowledge on German newspaper formats, we consulted only the news and business sections and excluded regional and local pages for Project A. Hence, our sample reflects primarily the political news coverage of the SZ and thus the political debates within the national arena. For papers in Project B, we excluded the sports sections, leisure sections, letters to the editor, commercials and (for national papers) regional sections. Second, coders read all remaining sections and selected *all* articles, as soon as a reference was made to unemployment, irrespective of whether unemployment was the main story, a secondary topic or even an incidental reference. Articles on all these topics were included into the sample. The number of articles sampled per issues was therefore rather high. Interrelations between various policy or issue fields and discourse arenas can thus be captured within the data base.

6.5 Number of Events or Claim per Article

For Perspective 1, sampling is now finished. For Perspectives 2 and 3, researchers additionally have to decide which events and claims to sample from selected articles. This depends on the particular research object, because we could either be interested in one particular type of event (e.g., protests) or actor (e.g., the unemployed) or in entire contention or discourse (i.e., all events and actors). In both our research projects we sampled *all* claims stated in selected articles (for a definition see Giugni/Statham, 2002). However, researchers need to define first what they define as an event or claims, and what kind of information has to be included at least in the newspaper to treat them as such. The reason for this is the fact that newspapers center their reports on some outstanding 'news', relegating other events or claims to secondary position (comments, reactions, contextual or circumstancial information etc.). In some cases, it is difficult to extract enough information from newspaper articles in regard to all cited or insinuated events and claims.

6.6 Triangulating Different Sample Types

In summary, we created two different samples. For Project B, we analyzed a wide array of newspapers longitudinally. This implied an extensive sampling strategy, forcing us to reduce the amount of articles per newspaper and the number of issues per year. This strategy is the best option to study the argumentative, ideational or ideological structure of public discourses. The topic of interest is long-term changes, which can be grasped only when analysing a longer pe-

riod of time and a bigger number of newspapers representing various positions and arenas. The sample of Project B thus resides on the assumption that arguments, ideas or ideologies dominating a public debate at a certain point of time, will probably be traceable even in every single article. For this reason we reduced the number of texts, trusting to find at least traces of these ideas there as well.

In Project B we analysed a number of newspapers for each point in time. Thus, we are able to survey a broader scope of the discourse by respecting different newspapers with their particular foci, topics and political orientations. While specific issue debates cannot be constructed adequately by operating with a bigger number of newspapers, the sample is very inclusive in terms of actors, arguments and ideas. In principle, this sampling strategy is best suited for an investigation of broader discursive fields because we are able to reconstruct a wider arena of claims-making actors with their particular concerns, agendas and opinions. However, our findings indicate that mass-mediated discourses are highly selective and thus centre on a restricted number of actors, even if we consider very different newspapers. A more encompassing sampling strategy would therefore need to include other media organs (TV, radio, internet etc.), too.

In Project A we focused on only one newspaper within a shorter period of time. This allows us to analyse the debate very intensively by sampling a big number of articles per issue and of issues per year. The decision to use one newspaper reduces the spectrum of news coverage and increases thus the data production bias. However, this might not be the primary problem from the particular research question's point of view. In fact, this kind of sample allows for an in depth analysis of issue specific debates by providing information on the semantic, syntactic and pragmatic structure of public reasoning. Moreover, we attained a more comprehensive picture of these debates by assembling articles on neighbouring issue fields (e.g., unemployment and labour markets, fiscal policies, social welfare, technological advances, international competition) and by tracing back the interrelations between various discourses (e.g., scientific, political, economic or administrative debates). Additionally, we are able to reconstruct the thread of events and claims more closely. The interactivity of public discourses cannot be captured in detail, yet, we get a sense of the interactive structure of public contentions. Hence, the biased picture of public discourses, which emanates from the analysis of the SZ, is a problem, though not necessarily a decisive one, as long as we opt consciously to analyse a particular spectrum of the public arena, e. g. the core policy domain of the most dominant political actors, onto which all prestige newspapers tend to centre anyway.

The decision to blend two sampling strategies was motivated, first, by the attempt to generate an enlarged data set fusing the strengths of each of the two procedures. This allows us to study public discourses on unemployment more comprehensively, i.e., by amalgamating a longitudinal study of discourse arenas with a cross-sectional analysis of issue debates and fields. Second, each sample has its drawbacks: If many articles from one newspaper are analysed, the selected newspaper might mirror the overall public discourse well; if few articles from many newspapers are analysed, chosen articles might not be typical of the respective newspapers. We thus attempted to minimize the biases of each procedure by triangulating the samples. On the one hand, project A provides useful information for determining how much data is necessary for a longitudinal analysis of discourse developments in Project B. Indeed, sample A suggests that broad political strands of argumentation can be uncovered within a very reduced number of articles. Yet, this is only possible if researchers apply interpretative methods of analysis. These results can then be fed into a more quantitative content analysis later on, but the first step is decisive when sampling is concerned.

On the other hand, project B can help to determine whether the SZ, which provides the empirical basis of Project A, is typical of German discourse on unemployment. Our findings indicate that this is indeed the case. Most newspapers privilege the core policy actors to an exceptionally high extent. Newspapers differ primarily in grades of selectivity and commentaries.

Unemployment initiatives are excluded from *all* selected print media. Apart from this, the SZ is centred on the two leading political parties and social partners, experts and think tanks. In comparison, unions are overrepresented, and liberal democrats underrepresented slightly. Yet, as Project A's data base is very large (N=2700), we have enough information about all pertinent policy actors, even some welfare organizations.

7. CONCLUSION

Process-generated data can be a valuable alternative to survey data and interviews, if the latter are not available. For some research questions process-generated data is actually better suited. In order to sample process-generated data, researchers first have to define the target population and cases. This definition depends on the research question. In contrast to surveys, several contrasting definitions of the target population are possible for the same data type. The reason for this is the multiperspectivity of process-generated data. They can be read in different ways, drawing different kinds of information from them. For example, newspapers can be used as data source for at least three different types of information: journalists' messages, mass mediated public discourse and historical facts. Researchers can draw most information on the first and least on the last question from newspapers. Because of this multiperspectivity and ambivalence, it is recommendable to analyse process-generated data first in detail using interpretative methods (Hanawalt, 1991). Only afterwards can researchers decide if quantification makes sense for the particular data type.

Process-generated data is also usually biased. If and how data is biased during data production, depends first on the definition of the target population: If the newspaper or journalists' messages are of interest, data production bias is usually no problem. For all other perspectives on the data, the production of news generates a bias that has to be taken into consideration. Here, we argued that the production bias is determined by the mass media's purposes, formats and social contexts. Bias varies over time, locally and content-wise. Process-generated data may additionally suffer from a data selection bias: Data may be destroyed, decayed or stored inaccessibly. For newspapers, data selection is usually no problem, although researchers need to know the social context well in order to find respective data sources. For other data types, data selection might produce additional problems. Due to biases, researchers usually cannot apply inferential statistics when using process-generated data. Instead, it is sensible to use alternative sampling strategies. In addition, data should be triangulated with other data sources in order to assess how the bias works.

Only after defining the target population and assessing bias can researchers sample process-generated data. In contrast to survey data, it is almost impossible to give general rules for sampling process-generated data. For newspapers, researchers have to decide which media organs, sections or articles they want to choose. Thus, they can use multi-staged sampling methods. For a given sample size, researchers face a trade-off along five dimensions: (a) the number of newspapers; (b) the number of years; (c) the number of issues per year and newspaper; and (d) the number of articles per issue. Which strategy is most suitable, depends on the particular research question. However, we argue that a number of rules and recommendations can be formulated when departing from different types of research questions and objects.

8. REFERENCES

Abbott, Andrew (1995). Things of Boundaries. *Social Research* 62. 857-882. Reprinted in: Andrew Abbott (ed.) (2001): Time Matters. On Theory and Method. Chicago: University of Chicago Press. 261-279.

Altmeppen, Klaus-Dieter (2000). Funktionale Autonomie und organisationale Abhängigkeit. Inter-Relationen von Journalismus und Ökonomie. In: Löffelholz, Martin (ed.) (2000): *Theorien des Journalismus*. Wiesbaden: Westdeutscher Verlag, S. 225-239.

Baum, Annerose/Lahusen, Christian (2004). *Germany. National Report on claims-making data.* Report of the Project: *The Contentious Politics of Unemployment in Europe. Political Claim-making, Policy Deliberation and Exclusion from the Labour Market*. http//ics.leeds.ac.uk/eurpolcom/unempol/papers.cfm.

Baur, Nina (2001). *Soziologische und ökonomische Theorien der Erwerbsarbeit. Eine Einführung*. Frankfurt a. M. / New York: Campus.

Baur, Nina (2004). Wo liegen die Grenzen quantitativer Längsschnittsanalysen? *Bamberger Beiträge zur empirischen Sozialforschung*.

Behnke, Joachim / Behnke, Nathalie / Baur, Nina (2004). *Empirische Methoden der Politikwissenschaft*. Paderborn: Ferdinand Schöningh.

Blossfeld, Hans-Peter / Hamerle, Alfred / Mayer, Karl Ulrich (1986). *Ereignisanalyse*. Frankfurt a. M.

Bow, James (1980). The Time's Financial Markets Column in the Period Around the 1929 Crash. *Journalism Quarterly*. Volume 57, No. 2, S. 447-450.

Bright, Robert / Coburn, Elaine / Faye, Julie / Gafijczuk, Derek / Hollander, Karen / Jung, Janny / Syrmbos, Helen (1999). Mainstream and Marginal Newspaper Coverage of the 1995 Quebec Referendum. An Inquiry into the Functioning of the Canadian Public Sphere. *CRSA / RCSA*. 36.3. 313-330.

Brookes, R. (2004). The media representation of public opinion: British television news coverage of the 2001 general election. *Media, Culture & Society*, Vol. 26, N. 1, 63-81.

Castles, Stephen/ Kosack, Godula (1985). *Immigrant Workers and Class Structure in Western Europe*. 2nd Edition. London et al.

Chiricos, Ted / Padgett, Kathy / Gertz, Marc (2000). Fear, TV news, and the reality of crime. *Criminology*, Vol. 38, N. 3, 755-787.

Cochran, William G. (1972). *Stichprobenverfahren*. Berlin / New York: Walter de Gruyter.

Creswell, John W. (1998). *Qualitative Inquiry and Research Design. Choosing Among Five Traditions*. Thousand Oaks / London / New Delhi: Sage.

Esping-Andersen, Gøsta (1990). *The Three Worlds of Welfare Capitalism*. Cambridge (UK) / Oxford (UK): Polity / Blackwell.

Flick, Uwe (2000). Triangulation in der qualitativen Forschung. In: Flick, Uwe / Kardoff, Ernst von / Steinke, Ines (ed.) (2000): *Qualitative Sozialforschung. Ein Handbuch*. Reinbek: Rowohlt Taschenbuch Verlag. 309-318.

Franzosi, Roberto (1987). The Press as a Source of Socio-Historical Data: Issues in the Methodology of Data Collection from Newspapers. *Historical Methods, 20*, 5-16.

Galtung, J. / Ruge, M. (1973). Structuring and selecting news. In S. Cohen & J. Young (Eds.), *The Manufacture of News: Social Problems, Deviance and the Mass Media* (pp. 62-72). London: Constable.

German Information Center (1995). *Citizenship and Naturalisation.* http://langlab.uta.edu/GERM/GIC/focus/FOCUS_95_2_January.txt (February 1995; 950 Third Avenue. New York).

Gesterkamp, Thomas (1993). So selbstverständlich und doch so unbekannt. Die Arbeitswelt in den Medien. *Medium*. Volume 23, No.4, 4-6.

Gigerenzer, Gerd (1999). Über den mechanischen Umgang mit statistischen Methoden. In: Roth, Erwin / Holling, Heinz (ed.) (1999): *Sozialwissenschaftliche Methoden. Lehr- und Handbuch für Forschung und Praxis*. 5.Auflage. München / Wien: R. Oldenbourg. S. 607-618.

Giugni, Marco / Statham, Paul (2002). *The Contentious Politics of Unemployment in Europe. Political Claim-Making, Policy Deliberation and Exclusion from the Labor Market. Codebook Workpackage 1: Political Claim-Making in the Public Domain*. Reihe: European Political Communication Working Paper Series. Volume 2/02. Leeds: http://ics.leeds.ac.uk/eurpolcom/unempol/papers.cfm

Hanawalt, Barbara A. (1991). The Voices and Audiences of Social History Records. *Social Science History*. Volume 15. No. 2. 159-175.

Hocke, Peter (1998). Determining the Selection Bias in Local and National Newspaper Reports on Protest Events. In: Rucht, Dieter / Koopmans, Ruud / Neidhardt, Friedhelm (ed.) (1998): *Acts of Dissent*. Berlin: Edition Sigma.

Hoffmann, Jochen / Sarcinelli, Ulrich (1999). Politische Wirkungen der Medien. In: Wilke, Jürgen (ed.) (1999): *Mediengeschichte der Bundesrepublik Deutschland. Bonn: Bundeszentrale für politische Bildung*. S. 720-748.

Keppler, Angela (2000): Medien und Kommunikationssoziologie. Verschränkte Gegenwarten. Die Untersuchung kultureller Transformationen. *Soziologische Revue*. Volume 23, Sonderheft 5, 140-153.

Kepplinger, Hans Matthias (1994). Publizistische Konflikte. Begriffe, Ansätze, Ergebnisse. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*. Sonderheft 24. Vol. 46. 214-233.

Küng, Lucy (2001). The Internet's impact on incumbent media firms. A management perspective. *Medien & Kommunikationswissenschaft*. Volume 49, No. 2, 218-226.

Lahusen, Christian/Baur, Nina (2005). *Do Media Discourses Influence.Opinions? Connecting Newspaper Discourse on Unemployment with Attitude Change on Social Security*. Paper Presented on the RC 33 Sixth International Conference on Social Science Methodology: Recent Developments and Applications in Social Research Methodology. 16 – 20. August 2004, Amsterdam. Paper on this CD-Rom.

Lerg, Winfried B. / Schmolke, Michael (1995). Die Zeitung als Quelle. Das Auswahlproblem. *Relation. Medien – Gesellschaft – Geschichte*. Volume 2. No. 1. 11-18.

MacDevitt, Michael (2003). In defense of autonomy. A critique of the public journalism critique. *Journal of Communication*, Vol. 53, N. 1, 155-165.

MacQuail, Denis (1992). *Media performance. Mass communication and the public interest.* London. Sage.

MacQuail, Denis (2003). *Mass Communication Theory*. London: Sage, 4th ed.

Mayer, Karl Ulrich (1998): Causality, Comparisons and Bad Practices in Empirical Social Research. A Comment on Stanley Lieberson's Chapter. In: Blossfeld, Hans-Peter / Prein, Gerald (ed.) (1998): *Rational Choice Theory and Large-Scale Data Analysis*. Boulder: Westview. S. 146-157.

McCarthy, John D. / McPhail, Clark / Smith, Jackie (1996). Images of Protest. Dimensions of Selection Bias in Media Coverage of Washington Demonstrations, 1982 and 1991. *American Sociological Review*. Volume 61. 478-499.

Müller, Albert (1996). Die Zeitung. Eine Quelle der Historischen Sozialwissenschaften. *Relation*. No. 1 3.1996. Vol. 3. 45-48.

Oliver, Pamela E. / Myers, Daniel J. (1998). How Events Enter the Public Sphere. Conflict, Location and Sponsorship in Local Newspaper Coverage of Public Events. *American Journal of Sociology*. Volume 105. 38-87.

Oliver, Pamela E. / Maney, Gregory M. (2000). Political Processes and Local Newspaper Coverage of Protest Events: From Selection Bias to Triadic Interactions. *American Journal of Sociology*. Volume 106. No. 2.

Ostner, Ilona (1995). Arm ohne Ehemann? Sozialpolitische Regulierung von Lebenschancen für Frauen im internationalen Vergleich. *Aus Politik und Zeitgeschichte.* B 36-37. 3-12.

Pietilä, Veikko (1992). Beyond the News Story: News as Discursive Composition. *European Journal of Communication*. Volume 7. No. 1. 37-68.

Pöttker, Horst (2002). Wann dürfen Journalisten Türken Türken nennen? Zu Aufgaben und Systematik der Berufsethik am Beispiel des Diskriminierungsverbots. *Publizistik. Vierteljahreshefte für Kommunikationsforschung*, Vol. 47, N. 3, 265-280.

Schmitt-Beck, Rüdiger / Pfetsch, Barbara (1994). Politische Akteure und die Medien der Massenkommunikation. Zur Generierung der Öffentlichkeit in Wahlkämpfen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*. Sonderheft 34. Volume 46. 106-138.

Schnell, Rainer (1986). *Missing-Data-Probleme in der empirischen Sozialforschung*. Inaugural-Dissertation zur Erlangung des akademischen Grades eines Doktors der Sozialwissenschaft an der Ruhr-Universität Bochum – Abteilung Sozialwissenschaft.

Schnell, Rainer (1997). *Nonresponse in Bevölkerungsumfragen. Ausmaß, Entwicklung und Ursachen*. Opladen: Leske + Budrich.

Schulz, Rüdiger (1999). Nutzung von Zeitungen und Zeitschriften. In: Wilke, Jürgen (ed.) (1999): *Mediengeschichte der Bundesrepublik Deutschland*. Bonn: Bundeszentrale für politische Bildung. S. 401-425.

Seale, Clive (1999). *The Quality of Qualitative Research.* London / Thousand Oaks / New Delhi: Sage.

Shoemaker, Pamela J. (1991). *Gatekeeping*. Newbury Park, Calif.: Sage.

Steinhage, Nikolei (2000). Zeitaggregation und Ereignisdaten. Eine Simulation zu den Auswirkungen der Prozeßzeitskalierung. *Globalife Working Paper* No. 3 / 2000. Fakultät für Soziologie an der Universität Bielefeld. Bielefeld.

Trappel, Josef / Meier, Werner A. / Schrapa, Klaus / Wölk, Michaela (2002). *Die gesellschaftlichen Folgen der Medienkonzentration. Veränderungen in den demokratischen und kulturellen Grundlagen der Gesellschaft.* Opladen: Leske und Budrich.

Tudor, Andrew (1992). Them and us. Story and stereotype in TV world cup coverage. *European journal of communication*, No. 3, Vol. 7, S. 391-413.

van den Berg, Harry / Smit, Johannes H. / van der Veer, Kees (1992). Contextualization in Newspaper Articles: A Sequential Analysis of Actors' Quotes on the PACTO Affair. *European Journal of Communication.* Vol. 7. 359-389.

van Dijk, Teun A. (1988). *News Analysis. Case Studies of International and National News in the Press.* Hillsdale, N.J.: Lawrence Erlbaum Ass.

van Dijk, Teun A. (ed.). (1997). *Discourse as Social Interaction. Discourse Studies: A Multidisciplinary Introduction.* London: Sage.

Wilke, Jürgen (ed.) (1999). *Mediengeschichte der Bundesrepublik Deutschland*. Bonn: Bundeszentrale für politische Bildung.

Wittenzellner, Helmut (ed.) (2000). *Internationalisierung der Medienindustrie. Entwicklung, Erfolgsfaktoren und Handlungsempfehlungen.* Stuttgart: LOG_X Verlag.

Wolf, Michael J. (1999). *The Entertainment economy. How mega-media forces are transforming our lives*. New York: Times Books.

Zakrzewski, Raimund H. (1995). Marketingforschung für eine Tageszeitung. Primär- und Sekundärerhebungen der Süddeutschen Zeitung. In: Böhme-Dürr, Karin / Graf, Gerhard (ed.) (1995): *Auf der Suche nach dem Publikum. Medienforschung für die Praxis*. Konstanz: UVK. S. 45-67.