TS •> Tijdschrift voor tijdschriftstudies ISSN 1386-5870

DOI: https://www.doi.org/10.18352/ts.343 TS •> #38, December 2015, p. 43-50.

Publisher: www.uopenjournals.org. Website: www.tijdschriftstudies.nl

Using Digital Archives in Quantitative Discourse Studies: Methodological Reflections

KOBIE VAN KRIEKEN

k.vankrieken@let.ru.nl

ABSTRACT

This methodological essay discusses the possibilities of using digital archives in quantitative discourse studies. I illustrate these possibilities by discussing a study in which the digital archive Delpher was used to build a relatively large corpus of newspaper narratives (*N*=300) in order to test hypotheses about the historical development of linguistic features associated with objective and subjective reporting. The large amount of data collected in digital archives like Delpher facilitates the construction of corpora for such hypothesis-driven studies. However, the collection of newspaper articles on Delpher in fact constitutes only a small, non-random and continuously changing selection of all available data. Due to these characteristics, the use of Delpher jeopardizes two core values of quantitative empirical research: the generalizability and the replicability of findings. Although these issues cannot be easily overcome, I argue that digital archives have the potential to broaden the methodological scope of discourse studies and increase the overall significance of the field.

KEYWORDS

Delpher, discourse studies, hypothesis-driven research, newspaper narrative

INTRODUCTION

In this methodological essay, I reflect on the use of digital archives in discourse studies. I focus on the use of Delpher as a resource to construct newspaper corpora with the purpose of quantitatively examining diachronic developments in journalistic discourse. Thus far, the vast majority of studies examining language use in historical newspapers have been highly qualitative and descriptive in nature.[1] Although these studies provide

[1] E.g., Donald Matheson, 'The Birth of News Discourse: Changes in News Language in British Newspapers, 1880-1930.' *Media, Culture & Society* 22:5, 2000, 557-573; Kevin Williams, 'Anglo-American Journalism: The Historical Development of Practice, Style and Form.' In Marcel Broersma (ed.), *Form and Style in Journalism: European Newspapers and the Representation of News 1880-2005*. Leuven/Paris/Dudley: Peeters 2007, 1-26.

valuable views on journalistic discourse in a historical context, they fall short of producing robust and generalizable results on historical developments in the linguistic features of newspaper articles. Quantitative examinations of such features remain scarce due to practical inconveniences, including the amounts of time and energy needed to construct large corpora using analogue archives.

Moreover, the quantitative studies which do attempt to systematically analyze journalistic discourse are often limited in terms of corpus size, time span covered, or complexity of the variables under scrutiny. Many of these studies focus on diachronic changes in objectivity, a concept which has been operationalized in many divergent ways. Stensaas, for instance, defines objectivity as 'a reportorial form' which 'contains only verifiable assertions, does not make claims to significance, and avoids statements of prediction, value, advocacy, or inductive generalizations without clear attribution to source'.[2] The unit of analysis in this study was the news article. This means that each article included in the corpus was coded either as an objective article or as a nonobjective article, using the criteria as mentioned in the above definition. In this approach, news articles with one, ten, or fifty 'statements of prediction' are all classified as nonobjective articles. This approach seems to disregard the possibility that the quantity as well as the specific types of objectivity markers may vary across news articles, and, accordingly, that news articles may vary in their degree of objectivity.

Other studies have operationalized objectivity as a discourse structure in which the most important and recent information is presented first, followed by less important and older information. This so-called inverted pyramid structure is often considered to be one of the hallmarks of objective journalism.[3] In a large-scale corpus analysis of 5,000 Australian newspaper articles published in 2007 and 2009, Johnston and Graham counted the number of articles with an inverted pyramid structure, the number of articles with a narrative structure, and the number of commentary articles.[4] A fourth category was distinguished to account for hybrid articles which combine an inverted pyramid structure with narrative elements. Results showed that the percentage of hybrid articles had increased somewhat between 2007 and 2009, whereas the percentage of narrative articles had decreased. In this study, too, the unit of analysis was the news article. Due to this broad categorization, it remains unclear exactly how objective the hybrid articles are and to what extent they differ in quantity and type of objectivity markers from inverted pyramid articles on the one hand and narrative articles on the other. In addition, the comparison between two years makes it difficult, if not impossible, to draw any conclusions on diachronic developments in the use of objective discourse structures.

[2] Harlan Stensaas, 'Development of the Objectivity Ethic in US Daily Newspapers.' *Journal of Mass Media Ethics* 2:1, 1986, 50-60 (53).

[3] David Mindich, *Just the Facts: How "Objectivity" Came to Define American Journalism*. New York & London: New York University Press 1998.

[4] Jane Johnston and Caroline Graham, 'The New, Old Journalism.' *Journalism Studies* 13:4, 2012, 517-533.

Finally, Høyer and Nossen examined diachronic changes in the 'stylistic features' of a large Norwegian newspaper.[5] They compared news articles published at four different points in time between 1950 and 2008 and concluded that journalists have become 'more visible in the text', which seems to indicate a change towards a less objective and a more subjective reporting style.[6] Unfortunately, however, their analysis of the journalist's visibility was restricted to the absence or presence of a byline revealing the journalist's name, contact details, and/or picture. As such, this study reveals more about diachronic developments in contextual rather than textual markers of journalistic objectivity and subjectivity.

In sum, these previous studies covered relatively short time spans and/or used broad coding categories which reveal little about the linguistic manifestations of objectivity in news discourse. Additional, systematic, yet fine-grained analyses of journalistic discourse are desirable to arrive at a more thorough understanding of developments in objective reporting. Digital archives facilitate the realization of such analyses.

CONSTRUCTING CORPORA USING DIGITAL ARCHIVES

In comparison to analogue archives, digital archives offer the possibility to build corpora in a relatively short period of time and thus facilitate large-scale linguistic analyses of journalistic discourse. Specifically, digital archives facilitate the construction of corpora with the purpose of testing hypotheses about diachronic changes in journalistic discourse. In this essay, I will illustrate these possibilities by discussing a study in which the digital archive Delpher was used to build a relatively large corpus of newspaper narratives (N=300) to test hypotheses about historical developments in the use and functions of speech and thought reports.[7]

BACKGROUND OF THE EXAMPLE STUDY

The use of speech and thought reports can increase both the subjectivity and the objectivity of a news narrative.[8] The relative dominance of these two functions varies, such that any given speech or thought report serves either a dominant subjective function or a dominant objective function. The following excerpt illustrates both functions.

---

[5] Svennik Høyer and Hedda Nossen, 'Revisions of the News Paradigm: Changes in Stylistic Features between 1950 and 2008 in the Journalism of Norway's Largest Newspaper.' *Journalism* 16:4, 2015, 536-552.

[6] *Ibid*. (536).

[7] Kobie van Krieken and José Sanders, 'Diachronic Changes in Forms and Functions of Reported Discourse in News Narratives.' *Journal of Pragmatics*, forthcoming.

[8] Kobie van Krieken, José Sanders and Hans Hoeken, 'Blended Viewpoints, Mediated Witnesses: A Cognitive Linguistic Approach to News Narratives.' In Barbara Dancygier, Wei-Lun Lu and Arie Verhagen (eds.), *Viewpoint and the Fabric of Meaning: Form and Use of Viewpoint Tools across Languages and Modalities*. Berlin: Mouton de Gruyter, forthcoming 2016.

> When Lieke was in the bedroom with dad, he [*her partner*] appeared with a hood he had torn off a rain coat and stood behind father Jan. He interpreted the look Lieke gave him as permission: (1) "Do it!" And she also yelled: (2) "I'll stay with you forever!"
> He put the hood over Jan's face and pulled it backward. (3) "Hold it tightly," Lieke said. Yesterday she confirmed: (4) "Then dad was gone pretty fast."[9]

In this excerpt, the first three quotations refer to what the news source was saying *during* the news events. Since the journalist was not present at these events, these quotations are unverifiable and therefore increase the subjectivity of the narrative. The fourth quotation, by contrast, refers to what this news source said *after* the events took place, at a trial at which the journalist presumably was present. This quotation is, in other words, verifiable and its dominant function is thus to increase the objectivity of the narrative.

Analyzing speech and thought reports in light of their dominant function should deepen our understanding of developments in journalistic discourse in terms of objectivity and subjectivity. The aim of our corpus study was therefore to examine diachronic developments in the use, forms, and functions of reported discourse in Dutch news narratives about murder, published between 1860 and 2009. Since Delpher only contains articles published until 1995, we used LexisNexis as a second archive to extract articles for the period between 1990 and 2009. In light of the theme of this special issue, the focus of this essay is restricted to the use of Delpher.

Based on the somewhat contradictory views expressed in previous studies, we hypothesized an increase in reported discourse with a dominant subjective function as well as an increase in reported discourse with a dominant objective function.[10] In the following sections I will refer to this study to illustrate some of the methodological considerations involved in the use of digital archives to construct corpora for hypothesis-driven research.

CORPUS SIZE

In conducting a quantitative discourse analysis, a proper corpus size is crucial to ensure representativeness of the results obtained. Determining what constitutes a "proper" size depends, among other things, on the type of linguistic features of interest and the distribution of those features in the specific text types of interest. Since the considerations involved in corpus size determination are similar for studies using analogue and digital archives, I will not go further into these considerations but refer to the guidelines provided by Biber.[11] In our study, we determined that a corpus size of 300 articles – divided into 15 periods of each 20 articles – would ensure representativeness, especially

[9] *De Telegraaf*, 24 July 2008. 'Moordpaar Belooft elkaar Eeuwige Trouw in Gevang; Vader met Capuchon en Kussen Gewurgd.'

[10] Kobie van Krieken and José Sanders forthcoming.

[11] Douglas Biber, 'Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation.' *Literary and Linguistic Computing* 5:4, 1990, 257-269; Douglas Biber, 'Representativeness in Corpus Design.' *Literary and Linguistic Computing* 8:4, 1993, 242-257.

since our unit of analysis was the sentence rather than the text. The total number of sentences in our corpus was 6,999.

SELECTION OF NEWSPAPERS

In selecting newspapers for inclusion, it is important to keep in mind that the population of newspapers and articles in digital archives is often not equal to the population of newspapers and articles in analogue archives. That is, digital archives usually do not contain all newspapers and articles available.[12] As of August 2015, for instance, Delpher contains eight million newspaper pages. Although this might seem an incredibly large number, it only represents about 10 per cent of all published newspaper pages. A sample of digital newspapers or articles might therefore be representative of the digital archive, but not of *all* newspapers and *all* articles. At a bare minimum, this incompleteness requires caution in interpreting and generalizing the results of a study. Moreover, it requires a careful selection procedure in the corpus construction phase.[13]

On a similar note, digital archives are constantly in the process of being updated and complemented with new materials. In April 2014 and in March 2015, tens of thousands of newspaper issues were added to Delpher. Although these stage-wise expansions benefit the generalizability of quantitative findings, they simultaneously jeopardize their replicability by creating the undesirable scenario in which search results are dependent upon the date(s) of the search. A given search of the archive might yield results that differ from the results of the exact same search conducted a couple of months earlier or later, which means that peer researchers are deprived of the possibility to conduct exact replications of a given study.

For our corpus, we selected newspapers which were digitized for longer periods of time. Unfortunately, however, none of the newspapers appeared and/or was digitized for the entire period between 1860 and 2009. Additional newspapers therefore had to be added for some of the periods. Our final corpus contained articles from 17 different newspapers, with articles from 4 to 7 different newspapers per decade. Included were local and national newspapers, newspapers with a conservative orientation and newspapers with a progressive orientation, and broadsheet newspapers and tabloid newspapers. And, of course, all newspapers were Dutch newspapers, since Delpher only contains Dutch language sources.

The advantage of including many different newspapers is that it creates variety and thereby the possibility to generate general views on historical developments in journalistic discourse. The variety advantage has a risky downside, however: historical developments that are found may in fact not be developments but merely differences between newspapers. For instance, if a given period includes articles from newspaper A but not from newspaper B, and the subsequent period includes articles from newspaper B

[12] See also Thomas Smits. 'TS Tools: Problems and Possibilities of Digital Newspaper and Periodical Archives.' *Tijdschrift voor Tijdschriftstudies/Journal for the Study of Periodical Media* 36, 2014, 139-146.

[13] See also Marcel Broersma, 'Nooit meer Bladeren? Digitale Krantenarchieven als Bron.' *Tijdschrift voor Mediageschiedenis* 14:2, 2012, 29-55.

but not from newspaper A, any differences between these periods might in fact reflect nothing more than differences between newspapers A and B. In selecting articles for our corpus, we therefore took care to provide as much overlap in newspapers between the decades as possible, so that each consecutive decade included articles from at least two of the same newspapers. Nevertheless, it is at all times recommended to perform analyses over the individual newspapers to control for possible differences between newspapers.

SELECTION OF ARTICLES

Random sampling of newspaper articles is the preferred method to construct a corpus for quantitative, hypothesis-driven research. This means that articles should be randomly selected from the population, such that each article has an equal chance of being included in the corpus. Random sampling techniques ensure that the results of a study can be generalized over the population. Above I noted that Delpher contains about 10 per cent of all available newspaper articles. These articles are not randomly selected from all available articles. Instead, a committee determines whether a newspaper is added to the digital archive based not only on pragmatic considerations but also on political, cultural, and religious criteria.[14] For example, newspapers which demonstrate 'an original articulation of a particular cultural or religious spirit' are more likely to be included on Delpher than newspapers which do not. Newspaper articles articulating a cultural or religious spirit are therefore likely to be overrepresented on Delpher. A random selection of Delpher articles is, in other words, not a random selection of all newspaper articles, which negatively affects the generality of the results of a study. This is especially problematic for discourse studies, since the ideological or political orientation of a newspaper is often reflected in its language use.[15] In assessing the implications of quantitative findings, researchers should therefore always bear in mind the selection criteria which guide the inclusion of newspaper issues in digital archives.

The use of non-random sampling techniques is generally discouraged in quantitative corpus analysis because the findings of a non-random sample cannot be generalized to the population. However, sometimes a non-random sample fits the purpose of a study best. This was the case in the example study, in which we aimed to examine diachronic developments in the specific genre of newspaper narratives about the specific topic of murder. Since of course not all newspaper articles are about murder and not all articles are narratives, we combined a judgment sampling technique with a quota sampling technique to select articles.[16] This means, first of all, that we searched for articles about murder and then judged for each article whether it was a narrative or not. Articles thus had to meet two criteria in order to be included: they had to describe a murder case and they had to provide chronological details about the events, implying

[14] The selection procedure is explained on the website of Delpher.

[15] E.g., Argiris Archakis and Villy Tsakona, 'Parliamentary Discourse in Newspaper Articles: The Integration of a Critical Approach to Media Discourse into a Literacy-based Language Teaching Programme.' *Journal of Language and Politics,* 8:3, 2009, 359-385.

[16] See Kimberly A. Neuendorf. *The Content Analysis Guidebook*. Thousand Oaks/London/New Delhi: Sage Publications 2002.

some degree of narrative reconstruction. Secondly, we searched per decade and selected for each decade the first twenty articles that met the criteria in order to ensure an equal distribution of narratives across the entire period between 1860 and 2009.

The great benefit of digital archives like Delpher is the possibility to browse newspapers and extract specific articles using search keys. When using search keys to extract articles, it is important to keep in mind that the translation of language from physical page to digital file via Optimal Character Recognition is not flawless. Old newspapers in particular suffer from incorrect character recognition. This is problematic for two reasons. First, the search results might include articles that are irrelevant to the study. In our study, for example, one of the search keys was *moord* ("murder"). The search results often displayed articles including words lexically similar to *moord* (e.g., *boord* "board/hem" or *Noord* "North") but not the word *moord* itself. Checking each article before including it in the corpus provides an easy solution to this problem. The second problem is more serious and comes without a solution: relevant articles might not be included in the search results. If, for instance, in a given article the word *moord* is falsely recognized as *boord*, that article will not appear in the search results. It is impossible to determine how many relevant articles are missed due to incorrect character recognition.[17] As such, this flaw imposes some serious restrictions on the generalizability of a study.

CONCLUDING REMARKS

The large number of newspaper articles collected on Delpher suggests completeness, but in fact constitutes a small, non-random and continuously changing selection of all available data. The use of digital archives like Delpher in quantitative discourse studies consequently jeopardizes two core values of empirical research: the generalizability and the replicability of findings. These issues cannot be easily overcome, but should by no means discourage researchers to use digital archives in quantitative corpus research. Meaningful results can be obtained as long as the incomplete, non-random and unstable nature of the archive is taken into account in (1) the selection of newspapers and articles for inclusion in the corpus and (2) the interpretation and generalization of the findings.

With due consideration of the pitfalls outlined in this essay, digital archives like Delpher are in fact promising tools in conducting quantitative discourse studies. One of the major benefits of digital archives is that they facilitate the construction of highly specific corpora for which non-random sampling techniques are required. In our example study, we constructed a corpus of articles in a specific genre (news narratives) about a specific topic (murder). Constructing this corpus using an analogue archive would have been a highly time-consuming, if not an impossible venture. Our corpus enabled us to test hypotheses about historical developments in journalistic discourse. Specifically, we were able to demonstrate that speech and thought reports have always been used to add both subjectivity and objectivity to news narratives, but that the dominant function of

[17] Marcel Broersma 2012.

these reports has shifted over time. Between 1860 and 1960, these reports were more often used to increase the subjectivity of a narrative than its objectivity. However, between 1960 and 2009, they were more often used to increase the objectivity of a narrative rather than its subjectivity.[18] Refined qualifications like these make an important contribution to previous findings on objective and subjective reporting in journalism and advance our understanding of developments in the pragmatic functions of language use in newspapers. Digital archives thus bear the potential to broaden the methodological scope of discourse studies and increase the overall significance of the field.

•> KOBIE VAN KRIEKEN *is a PhD candidate at the Centre for Language Studies at the Radboud University. Her research focuses on the form, function, and impact of news narratives*

[18] Kobie van Krieken and José Sanders forthcoming.