DE GRUYTER

EMPIRICAL MULTIMODALITY RESEARCH

METHODS, EVALUATIONS, IMPLICATIONS

Edited by Jana Pflaeging, Janina Wildfeuer, and John A. Bateman

Empirical Multimodality Research

Empirical Multimodality Research

Methods, Evaluations, Implications

Edited by Jana Pflaeging, Janina Wildfeuer and John A. Bateman

DE GRUYTER

ISBN 978-3-11-072491-2
e-ISBN (PDF) 978-3-11-072500-1
e-ISBN (EPUB) 978-3-11-072515-5

Library of Congress Control Number: 2021946552

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Printing and binding: CPI books GmbH, Leck

www.degruyter.com

Preface and Acknowledgements

The editors would like to thank, first and foremost, the authors of this book for their enthusiasm for our joint publication project, their excellent contributions, and all their efforts during the peer-reviewing process. We are also grateful for the insightful suggestions made by an anonymous reviewer in the post-submission phase. Without their dedication, this volume would not have been possible.

We also want to express our gratitude to the authors and all other participants for attending #BreMM19: *Fourth Bremen Conference on Multimodality: Empirical Inroads* on 25–27 September 2019 in Bremen, Germany. The exciting talks and engaging discussions held at the conference painted a lively and colorful picture of a research field 'on the move'. They demonstrated precisely how much can be gained when we keep aiming for (even) more objective research designs, (even) more systematic frameworks for reliable analyses, and (even) more careful checking of how well our findings match the questions we seek to answer.

Achieving a more solid empirical grounding of multimodality research requires following these core tenets of empirical research — objectivity, reliability, and validity. This allows for making new discoveries and for revisiting old ones. Finally, the conference also showed once more that conducting empirical research typically is a laborious task that demands pooling the expertise and resources from various disciplines, research fields, and institutions all around the globe. We gratefully acknowledge Carman Ng's invaluable help in organizing and running *BreMM19*, and her contributions to preparing its proceedings.

With the present volume, we not only seek to document the most recent advancements in empirical multimodality research as presented at *BreMM19*, but also look to capture the spirit of a general move towards empirical multimodality research. We hope that this book will contribute to this exciting development and inspire future empirical work.

The Bremen Conferences on Multimodality have become a staple within multimodality research. *BreMM14*, which took place in 2014, was dedicated to building bridges between various multimodality-ready disciplines. *BreMM15* continued with theoretical and methodological explorations in 2015, and paved the way for discussing and promoting multimodality's disciplinary status at *BreMM17*. Focussing on empirical inroads into multimodality research — the theme of *BreMM19* and the present volume — proved one of the next steps to take in advancing our research field with regard to its status as a discipline, and far beyond. In 2020/21, the tradition of regular meetings continues as *The Bremen-Groningen Online Workshops on Multimodality*. We hope to be able to keep this high level of continuity and fruitful exchange in one form or another and push multimodality's development further.

https://doi.org/10.1515/9783110725001-202

The *BreMM19* conference was a success also thanks to the generous financial support of the German Research Foundation (Deutsche Forschungsgemeinschaft — DFG) as well as the logistical support of the University of Bremen. We gratefully acknowledge their contributions to the conference.

Jana Pflaeging, Janina Wildfeuer, and John A. Bateman
Salzburg, Groningen, and Bremen, July 2021

Contents

Part I: Introduction

Part II: Charting Paths for Empirical Research: Theoretical and Methodological Reflections

Part III: Empirical Inroads: Case Studies and Results

Part I: **Introduction**

Jana Pflaeging, John A. Bateman, and Janina Wildfeuer

Empirical Multimodality Research: The State of Play

Abstract: Multimodality research has always shown a strong reliance on data. However, the field has primarily developed around more exploratory, descriptive, and interpretative work on smaller data sets — as suggested by results we present from a meta-study of contributions to three multimodality-close international journals (*Social Semiotics*, *Visual Communication*, *Multimodal Communication*). Framed by a discussion of the qualitative-quantitative dichotomy and a deliberately broad working definition of *empirical*, we argue that it is not sample size or quantitative methods alone that support a more solid empirical grounding of multimodality research, but rather an explicit orientation to just how theory and analysis make contact with data. To this end, we propose five quality criteria of empirical practice, that is, completing the empirical feedback loop: from theory to data and back, ensuring objectivity, reliability, and validity in research, and acknowledging the inherent tentativeness of results. We thereby seek to chart paths for an appropriate and productive application of various empirical methods to novel (and supposedly familiar) forms of meaning-making in order to further strengthen the development of theory and methods in multimodality, and to encourage an even more intense exchange among the diverse communities of multimodalists.

Keywords: empirical multimodality research, data, qualitative, quantitative, quality criteria

1 Introduction and the Aims of this Volume

As a research endeavor first and foremost borne out of the practical observation that all meaning making naturally involves a multitude of forms of expression, multimodality research has always been driven by data. When we look more at the kinds of explicitly ‘empirical’ work that have preoccupied multimodalists of many stripes over the past 25 years, however, it is fair to say that multimodality research locates itself mainly towards the smaller-scale and more qualitative poles of the empirical continuum. In many respects, this is understandable: driven by the challenges of engaging with new kinds of increasingly complex research objects, the field has developed around more exploratory, descriptive, and interpretative work undertaken with respect to smaller sets of data.

https://doi.org/10.1515/9783110725001-001

Nevertheless, as research objects continue to diversify, the interest of neighboring disciplines increases, and the field shows signs of becoming a stand-alone discipline in its own right (see Wildfeuer et al. 2019), the need for solid empirical grounding is also becoming ever more apparent. In fact, diverse scholars in multimodality have been pointing for some time, and often quite independently of one another, to the usefulness of 'large n' empirical investigations (see, e.g., Stöckl 1997; Bateman et al. 2004; Gu 2006; Carter & Adolphs 2008; Nakano & Rehm 2009; Bednarek 2015; Hiippala 2015; Pederson & Cohn 2016; Bezemer & Cowan 2021). Furthermore, scholars are increasingly stressing the shortcomings of exclusively broad-brushed orientations (e.g., Jewitt 2017; Kohrs 2018). Multimodality research as a whole thus seems poised at a particular point of development where exploratory studies can beneficially be complemented by further kinds of empirical work bringing the potential of a productive interaction across the entire empirical continuum into view.

This development requires careful consideration, however, and a certain 'hesitation to scale-up' is still common. There are several reasons for this, ranging on the one hand from work indeed being so experimental in nature that larger-scale studies might well be premature, to on the other hand, lack of knowledge and experience concerning just how such larger-scale studies might be conceived and conducted. In many institutions where approaches to multimodality are taught, methodologies for larger-scale studies are not prominent on the curriculum. Moreover, addressing this concern is not just a question of applying well-established techniques from elsewhere: there are also significant theoretical issues revolving around just how empirical methods can be productively applied to novel forms of meaning-making. It is often by no means clear how best to proceed and further methodological guidance — or: "greater rigor and investment of effort in developing robust conceptual frameworks and reliable methods" (Thomas 2019: 86) — is urgently required.

This challenging situation constitutes the overall context for the current volume. As larger multimodal corpora become available and computer-based tools are developed to assist the processing of greater quantities of multimodal data, scholars, often collaborating in teams across national, disciplinary, and methodological borders, increasingly seek a more solid empirical grounding for multimodality research. Some of these endeavors are documented in the contributions to this volume, where our goal has been to demonstrate a range of engagements with multimodality research that exhibits a strong empirical orientation. The various chapters in the book consequently include both example analyses where this is done and some of the methodological and theoretical concerns that such work raises. In addition, the need to make approaches to multimodally complex artifacts and performances more receptive to empirical grounding requires both more foundational work on methods and the adaption of relevant empirical methods from other research areas as well. Several chapters of the book address these concerns specifically.

One of the essential preconditions for making appropriate and productive contact with data is to complete the empirical 'feedback loop' *from theory to data and back to theory*. Without this loop, research in multimodality will continue to lack the solid empirical grounding that now appears necessary for progress. Establishing how this can be done in a methodologically appropriate fashion is, in our view, just as important as increasing sample sizes. It also legitimizes smaller-scale work whenever attention is paid to core tenets of conducting empirical research — that is, it is not sample size alone, or quantitative methods, that make the difference, but rather an explicit orientation to just how theory and analysis engage with data. Any turn to the empirical in multimodality research is therefore also in need of a more critical and thorough reflection on how this connecting of theory with data and vice versa is to be achieved. This is a dimension of empirical investigation that has long been taken for granted and is only recently beginning to receive the attention it requires within multimodality research. Deepening this discussion is then a main aim of this book, reflecting carefully on quality criteria for conducting empirical research with data-sensitive/-responsive concepts and frameworks suitable for the review and renewal of research practices and hypotheses.

For the purposes of framing the contributions collected in the volume and of positioning the view of empirical multimodality research we envisage more productively, we organize the discussion around the following three factors:

- **Methods.** The volume presents methods for investigating a broad variety of multimodal artefacts (corporate logos, advertisements, news texts, posters, films, video games) and performances (e.g., political TV interviews, face-to-face teaching, oral narrative), in which theoretical frameworks in multimodality research of rather different kinds are carefully applied to — typically — larger data sets.
- **Evaluations.** The case studies presented also support critical evaluations of existing theoretical and methodological frameworks. The book consequently includes several contributions that deal more exclusively with questions of moving from 'theory to empirical inroads' and which thereby evaluate current practices of applying theory to data.
- **Implications.** The contributions also reflect on the implications of their findings — be they of theoretical, methodological or analytical nature — and make concrete suggestions for the adaption and expansion of existing practices and the design of future research projects with an empirical slant.

We see these perspectives as offering particular insights on the process of conducting empirical research. However, before proceeding to the contributions themselves, we provide in this introduction a brief overarching characterization both of the nature of empirical research and its current state within the field of multimodality — we mentioned an 'empirical continuum' above, but just what does this entail? Providing more detail here will help anchor the various directions which the contributions to the volume illustrate against the backdrop of a growing orientation to empirical work in multimodality more broadly.

First, then, we address the traditional dichotomy of *qualitative* versus *quantitative* and relate it to the notion of *empirical*. This is important as a preliminary stage in bringing together formerly rather disjoint sets of methods adopted for multimodality research. Second, we ask just how *empirical* the field of multimodality has in fact already become. To answer this question, we present results generated in a quantitative study of publication output of three prominent international journals in the field: *Social Semiotics*, *Visual Communication*, and *Multimodal Communication*. Third, on the basis of our survey, we argue that promoting an 'empirical turn' in multimodality research is now justifiable and beneficial and, to encourage such work, we propose a list of five quality criteria that can be drawn on to shape empirical practice. Finally, we preview the contributions to the book from the perspectives outlined and draw out some broader implications for our understanding and practice of empirical multimodality research.

2 The Qualitative-Quantitative Dichotomy and the Notion of *Empirical*

Historical Roots. A look at some of the earliest endeavors in research reveals how scholars have always naturally conducted work that can be considered 'empirical' to generate knowledge. A more solid empirical grounding of multimodality research is thus by no means a daring new move, but rather a reorientation towards the close connection between a research interest in real-world phenomena and their in-depth study to generate answers. This realization also points to a common denominator in qualitative and quantitative approaches, whose *dichotomous* relationship results from historical convention rather than some inherent difference in nature.

To show this it is revealing to consider historical precedents. Up until the end of the Middle Ages, for example, a world view had been cultivated that was deeply entrenched by superstitious belief. The Scientific Revolution of the 16th century then opened up an eagerness to explore the natural world by means of controlled experiments and the invention of tools and instruments to pursue them (Kevles

1992: 12). This development saw the natural world as separate from the perceiving individual — a principal tenet of an epistemology glossed as *positivism* by Auguste Comte in the 19th century. In this view, an objective truth is made accessible through observation and testing of cause-effect relations, which in turn prepares the ground for unbiased explanation, generalization, and the establishment of universal laws (Bhattacharya 2008: n.pag.; Sousa 2014: 211; see also Kirk & Miller 1986: 14).

The Early Modern Period then witnessed the strengthening of alternative approaches to knowledge generation, such as *hermeneutics*, which is practiced by interpreting the written and spoken word, originally primarily biblical texts. A growing emphasis on intuitive processes of understanding lead Dilthey to postulate around 1900 what he perceived to be the main difference between the natural sciences and the humanities: while the former *explain* the world, the latter seek to *understand* it (Bühler 2003: 4). This epistemological development continued throughout the 20thcentury, perpetuating (supposedly) different practices of research and debate, as well as different attitudes and perceptions of one another (cf. Yanow & Schwartz-Shea 2015: xiii).

Research Paradigms. Research in the natural sciences is consequently interested in generating objective facts and revealing universal regularities. Typically, by convention, scholars investigate larger quantities of data that have been representatively sampled (to account for natural variation), and pursue approaches that involve controlled measurement procedures; their methodological orientation is thus *quantitative*. Equally interested in revealing regularities and patterns, the related branch of (empirical) sociology draws on the possibility of quantification as well, but pursues essentially *qualitative* approaches, and thus crosses the (supposed) border between quantitative and qualitative research (see Kromrey 2002). By tradition, the humanities in contrast show a strong leaning towards using introspection, interpretation, and subjective perspectives to achieve an in-depth understanding of the 'nature' or *quality* of things (Kirk & Miller 1986: 9).

Unfortunately, these traditions and conventions have tended to push humanistic scholarship into the position of a clear counter-player to natural scientific research, resulting in misconceptions such as "[i]f statistics and 'large n' studies [...] were to be understood as quantitative analysis, then 'small n' studies using nonstatistical methods [...] must be 'qualitative' analysis" (Yanow & Schwartz-Shea 2015: xiii; see also Riesenhuber 2009: 7). Such assumptions about sample sizes are inherently problematic because they suggest that qualitative humanistic research must conform to just those validity standards adopted for 'objective' quantitative natural science research for that scholarship to be considered worthwhile (cf. Bollnow 1974: 1; Sousa 2014: 212). In a similar fashion, some humanities scholars come to the counter-view that “the search for patterns, regularities, or laws has no place in the Humanities” (van Peer et al. 2007: 7).

Since the mid-20th century, such ‘two-fold taxonomies’ (Yanow & Schwartz-Shea 2015: xiii) of *quantitative* vs. *qualitative* and *natural science* vs. *humanities* have been subject to increasing critique. This is not least documented in C.P. Snow’s well-known Rede lecture of 1959, in which he describes two separate ‘cultures’ of scholars, the natural scientists and the ‘literary intellectuals’ (Snow 2001 [1959]: 2–4). While being “comparable in intelligence”, Snow had noticed that they had “almost ceased to communicate at all” and urged his audience to understand the “practical and intellectual and creative loss” because “we are letting some of our best chances [for discovery, JP/JB/JW] go by default” (Snow 2001 [1959]: 2–4, 11, 16). More recently, a growing appreciation has developed of the productive synergies that more integrative and collaborative approaches can support (Brannen 1992). In this context, scholars of ‘both trades’ engage in extensive discussions about the fallacies of their own paradigms. In particular, quantitative research, and the positivist or reductionist paradigms it has traditionally been associated with (Bollnow 1974: 2), is now commonly described as being equally based on theorization, and its questions and interpretations as socially derived (Bhattacharya 2008). Qualitative research, in turn, has been accused of relying too much on intuition and speculative reasoning (Sampson 2002: 2; van Peer et al. 2007: 7 both in reference to linguistics), and of cultivating a habit of generalizing their arguments despite a noticeable ‘gap in evidence’ (Piper 2016: 5–6, in reference to cultural studies).

Many of these broader social currents play out in microcosm in linguistics, which itself also has a considerable tradition in conducting empirical research dating back to the 18th century. Sub-fields such as computational and corpus linguistics, applied linguistics, phonetics, psycho- or sociolinguistics (and much interdisciplinary work) have long cultivated the use of empirical methods, with a pronounced emphasis on quantitative methods and experimentation (cf. Wasow & Arnold 2005: 1485). While scholars have acknowledged that linguistics “straddles the humanities/science borderline” (Sampson 2005: 17) and that there are valid areas of linguistics where the empirical quantitative methods do not apply (Sampson 2005: 17), they have also urged that there is still much room for further productive development bridging these perspectives (Sampson 2002: 1).

Consequently, scholars in linguistics and beyond have been reflecting upon what original contributions to knowledge generation they might make. Researchers with a leaning towards qualitative approaches, for instance, have begun to acknowledge and embrace the particular contributions made possible by their viewpoints. Arguments are made for their adequacy in generating valid scientific knowledge (Sousa 2014: 212) even if this challenges traditional notions of ‘truth’ and ‘evidence’ (Bhattacharya 2008). Although there is still a “need for clear criteria governing

… monitoring, rigour, and quality assessment" (Sousa 2014: 212), a more explicit legitimization of qualitative approaches grants them a secure position in more complex research procedures, for instance when developing hypotheses, which are then made testable through quantitative research (Riesenhuber 2009: 6, 7).

***Empirical* — A First Approximation.** These considerations make any subscription to particular research paradigms a matter of *more/less* rather than *either/or*. Indeed, a raised awareness of methodological diversity within both paradigms (Benoit & Holbert 2008: 622) blurs the traditional distinction between *quantitative* and *qualitative* forms of inquiry and has revealed a core interest that both the humanities and the natural sciences share: conducting *empirical research*.

In the widest sense, then, *empirical research* simply means seeking to answer research questions about real-world phenomena by means of studying intersubjectively observable data (be they real/authentic, manipulated in experimental settings, or even intuition-deduced), whose results are utilized to reassess previous knowledge structures and associated hypotheses (cf. Bateman & Hiippala, this volume). The question of how to accomplish the move from theoretical/hypothetical assumptions to data description and back, therefore, lies at the heart of empirical research, and so needs to be considered and reported on thoroughly. Contrary to the prevalent but misleading assumption introduced above that has tended to label methods according to the size of data samples, this also means that if the connection between theory and data is sound, the label *empirical* is not solely reserved for quantitative 'large n' studies and can apply equally to 'more qualitative' work, even when smaller in scale (cf. Benoit & Holbert 2008: 615).

This notion of empirical research will be expanded on in Section 4.1 below specifically for the multimodality case. This will allow us to set up a view of empirical multimodality research that is part of, and contributes to, other branches and directions of multimodality rather than being a disjoint 'school of thought'. This then furthers our main aims of encouraging productive dialogue and exchange between multimodality approaches that are small-scale, qualitative, and perhaps exploratory on the one hand, and approaches that are larger-scale with quantitative support on the other.

3 The State of Empirical Work in Multimodality Research

When multimodality research slowly came into its own back in the 1990s, we already find research that explicitly labels its approach or the data 'multimodal', i.e., as being characterized by a particular combination of modes, while also adopting explicitly ‘empirical’ orientations. There are, for example, clear overlaps in questions and, increasingly, in methods from work within human-computer interaction (HCI), multimodal document design, multimodal interaction and extended conversation analytic perspectives.

Among these approaches there is already a long tradition of applying diverse ranges of empirical methods. Such methods include eye-tracking (e.g., Bold & Herranz 1992; Thorisson et al. 1992; Koons et al. 1993), behavioral user studies (e.g., Giard & Peronnet 1999), and later also neurocognitive studies. There has also been work aimed at providing large-scale corpus analyses across genres or from a diachronic perspective that generally adopts a more quantitatively oriented, data-driven perspective. Now this form of empirical work is increasingly overlapping with approaches in which transcription has always been a substantial first step in analysis, but now commonly extended to the description of ‘additional’ interactional resources, such as gaze (e.g., Goodwin 1980) or gestures (e.g., Streeck 1983). Many of the challenges here are consequently common across both multimodal corpus linguistics and multimodal transcription (Thibault 2000; Norris 2002; Baldry & Thibault 2005, 2006). Many of the investigations in these contexts present the most detailed approaches to the combination of modes to date and so constitute indispensable resources for future empirical research.

30 years after the beginnings of such empirical work, we wanted to probe the question of *how empirical* multimodality research has in general become. For this, we undertook a systematic overview of empirical work as documented in contributions to several international journals clearly devoted to the study of multimodality. Our goal was to see if there has been any change over the past three decades concerning how articles present themselves along the qualitative-quantitative continuum. To obtain a view of a field and to get us closest to a blueprint of the scholarly debate (see Engels et al. 2018: 594–595), international journals offered a suitable object of study, enjoying quality control, a much wider distribution than monographs or handbooks, and easy accessibility.

The data for our survey was consequently sampled from three international journals that we take to be prominent for the broad ‘communities’ engaging in multimodality research: *Social Semiotics* (which commenced publishing in 1991), *Visual Communication* (2002–), and *Multimodal Communication* (2014–). For purposes of comparison across journals, we limited the articles from *Social Semiotics* to the time after 2000. Publications were then considered up to and including 2020. All data was gathered from the journals’ respective online archives and search engines. Since the total number of contributions did not appear to be directly queryable, we used a search for the term ‘communication’ as a proxy as we expected this to occur in most contributions. The resulting totals for *Social Semiotics*, *Visual Communication*, and *Multimodal Communication* were 682, 461, and 86 respectively. All references to total numbers of publications in the following draw on these values.

Each of the journals showed a different pattern concerning the number of published articles per year. *Social Semiotics* showed a slight but steady increase up until 2015 (from around 20 a year in 2000 to around 35 in 2015), while *Visual Communication* remained approximately constant at 25 over the same period. Then, after 2015, both journals showed a marked increase due to the transition to 'online first' publishing practices and backlogs in the publication queue being processed (with *Social Semiotics* reaching 60-80 articles in 2019-2020 and *Visual Communication* over 45). The trend for *Multimodal Communication* for the time period from its original appearance was quite different, however, with a slight decrease in published articles starting at around 18 per year and ending at around 10; the number is now increasing again. Our specific goal for this chapter was then to explore to what extent articles have been *explicitly identifying themselves* as more or less empirical in orientation. Subsequently, on the basis of this selection, we investigated the distribution of methods and sizes of data sets among those articles.

The first step in this procedure was to find articles that (even implicitly) 'self-identified' as being empirical in the loose sense of showing a concern with data. For this, we retrieved all papers that contained occurrences of keywords that we took to be particularly likely to indicate an empirical orientation. The keywords used were 'empirical', 'statistical', and 'calculate', each with morphological variations for tense, adverbial usage, etc. accounted for. Again for current purposes, we restricted hits to those papers where the keywords in any morphological form occurred in the main bodies of the articles, excluding occurrences only in the abstracts or the references.

Each article that was logged matching these search criteria was examined to rule out cases where the keywords had been used but the article did not, in fact, exhibit any empirical orientation. For the articles remaining we recorded the journal in which they appeared, the year of publication, the numbers of authors, and three further categories identifying the kinds of empirical work done. These were: (a) the type of empirical methods employed, either 'qualitative', 'quantitative', or 'mixed' (i.e., triangulating qualitative and quantitative methods), (b) the size of the data sets employed, coded as 'small', 'medium', or 'large', and, if applicable, (c) the types of statistical procedures employed, i.e. simple 'counting', 'descriptive', or 'inferential'. We give further details on the criteria used for these categories in our results below. The data were first collected in an Excel spreadsheet, with rows for each article and columns for each coding category, and then imported into R and R Studio (R Core Team 2016) for actual processing and visualization; graphing

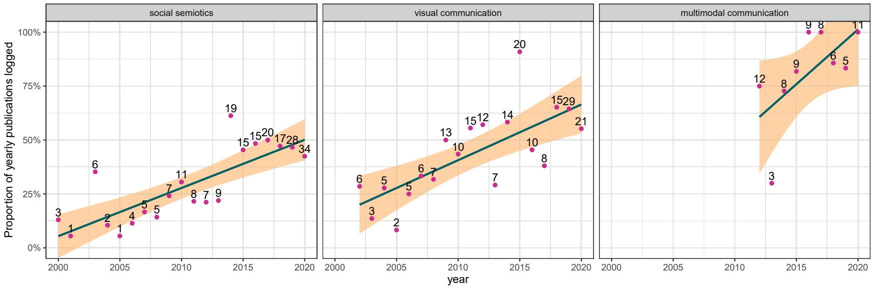


Fig. 1: Proportion of articles logged as broadly 'empirical' in the sense of being concerned with data. Each graph shows the number of articles logged per year for each journal and a fitted linear trendline with standard error indicated by shading (graphed with R 'ggplot2'). The trendline simply places a straight line approximating the data (see Bateman & Hiippala, this volume) to show broad relationships between proportions of empirical papers and years.

and visualization here is done with the R package 'ggplot2' throughout (Wickham 2016).

An initial question was to compare the number of articles retrieved and classified as empirical to the overall number of publications for the journal for each year. The purpose of this was to see whether the proportion of articles self-identifying as empirical had undergone any changes over time. The results are shown in Figure 1, which sets out for each year and for each journal the number of articles judged to be empirical in orientation expressed as a percentage of the total number of articles that appeared in that year. The results for all three journals seem to indicate that it is becoming more common for articles to explicitly frame their work as engaging with data and data analysis. For both *Social Semiotics* and *Visual Communication*, however, the earliest values are surprisingly low and so this may indicate that the articles before 2005 should be examined more carefully to see if they are formulating their engagement with data using words other than our adopted keywords. In the case of *Multimodal Communication*, we see, in sharp contrast, that the number of articles retrieved is a very high proportion of the total journal output, although again showing a marked upward trend over time.

We then turned to the kinds of empirical methods and the scale of the data employed in the articles retrieved and classified as empirical. The results of this part of our study are visualised in Figure 2. In this diagram, the first row sets out the proportions of logged papers that were classified according to the empirical approach adopted: qualitative, quantitative, or mixed. Each bar in the graphs shows how the logged papers divide up over those categories for that year — when the graphs are all one color, then there was only type of approach used in that year; when there are two colors, then the size of the colored regions shows the respective

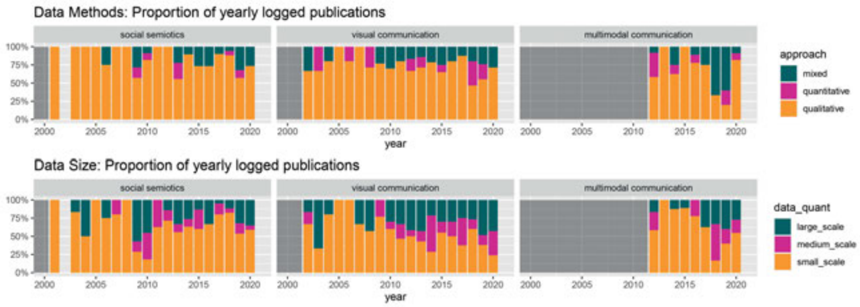


Fig. 2: Distribution of empirical methods and scale of data sets per year for each journal expressed as a proportion of the total number of articles tagged as broadly empirical each year. The grayed out areas are years where the journal in question did not appear or which lay outside our dataset. (Graphed with R 'ggplot2'.)

proportions out of the total logged papers for that year; and so on. We can see, therefore, that for the rightmost graph in the top row, for the journal *Multimodal Communication*, the proportions change quite dramatically over the years sampled, with the papers using mixed methods taking large proportions in two of the years towards the end of the sample. The situation for *Visual Communication*, shown in the middle of the row, is more evenly distributed with the proportions for mixed methods remaining broadly the same from around 2008 onward. Nevertheless, there is an increase in the proportion of quantitative papers as well. The lower row of the diagram is read in a similar way, but in this case the proportions shown are for the size of the data sets analyzed, divided into small, medium, and large. For current purposes we set the cut-off values for these categories as 'small' being less than 20 analysis items, 'medium' between 21 and 60, and 'large' as more than 60.

Here we can see that for the majority of the time sampled, the vast majority of articles were qualitative and small scale. Particularly for *Visual Communication*, however, the proportion of larger scale studies has increased to a considerable degree, with the proportion of small scale studies falling below 50% of the empirical articles from 2015 onward.[1]

[1] A somewhat similar study also partly focusing on the methods employed in studies presented in *Visual Communication* between 2002 and 2019 has been conducted by Thomson (2021). The results of the study report that the journal published 276 articles using 23 separate methods with "about 89 percent of these […] empirical and about 11 percent […] non-empirical". 'Empirical' here mostly refers to the use of data in any form, and the author mentions both "discourse analyses, textual analyses, content analyses, historical approaches and case studies" as covering 65.15% as well as "interviews, experiments, observations, surveys, autoethnographic approaches, photo

A similar pattern can be seen for *Multimodal Communication*, although with considerable variation. *Social Semiotics* has stayed predominantly small scale and qualitative throughout — showing if anything a slight increase in small-scale studies over time, although this impression may be artificially induced by the oddly high number of large- and medium-scale articles appearing in 2009–2010. It is interesting to note for all the journals both that the proportion of articles that adopt ‘mixed’ methods is generally far higher and more frequent than those simply reporting on quantitative results and also that larger scale studies are more prominent than ‘medium’ scale studies. The reasons for this would need further study, but it may be influenced by researchers, if they are using empirical methods at all, attempting to increase the size of their data sets. This would clearly be in the spirit of the move towards more empirical work that we are promoting here.

Finally, we investigated whether or to what extent there had been a change in the kinds of statistical methods employed. For this we contrasted articles where the quantitative treatment of the data included simple counts of items, where it included basic descriptive statistics such as means and standard deviations, and where it included standard inferential statistics, such as tests of significance of various kinds. We made a general distinction between counting and descriptive statistics, since even the most basic engagements with data may indicate how many cases were being examined without considering further quantitative properties. In addition, we were quite broad in our interpretation of ‘inferential’, including cases where, for example, corpus annotation accuracy had been verified with inter-coder reliability tests, and so on. The three-way distinction can therefore best be seen as a general indication of the sophistication of the statistical methods employed; finer-grained characterizations could certainly be pursued in the future. The results of our current classification are visualized in Figure 3; this shows the distributions in a slightly different way to that of the preceding graphs. In this figure, the three graphs show the breakdown of logged papers according to the selected categories of types of ‘statistical’ approaches similarly to before, but now they are not made to sum to 100% because we simply omit the remaining articles where no statistical methods were identified. These articles are not shown in the counts of the graphs in order to leave the pattern clearer among those articles that *did* use some forms

voice approaches, participatory methods, and the focus group and Q-Method” as the remaining 34.85% of empirical methods. 10.49% of all articles that were annotated as ‘empirical’ were seen as quantitative approaches (see Thomson 2021). Focusing on geographical information as well as the visuals used in the papers, the study did not address trends or developments in the use of specific methods over the years. However, it becomes clear that, similar to our own study, the number of more quantitatively oriented approaches, such as experiments and participatory methods, is rather small (approx. 10%) in comparison to more qualitatively oriented approaches.

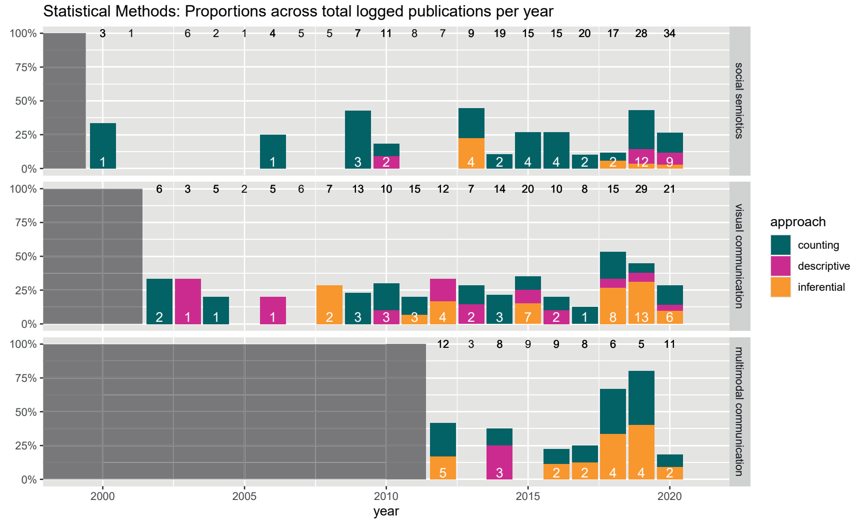


Fig. 3: Distribution of statistical methods per year for each journal, expressed as a proportion of the total number of articles logged as broadly empirical per year. The numbers at the bases of each column show the total number of quantitative papers per year. The numbers above each column show the total number of logged papers per year as were visible in Figure 1 above. The grayed out areas are years either where the journal in question did not appear or which lay outside our sample period. (Again all graphs were produced with R ‘ggplot2’.)

of measurement. The figure then gives a better impression of the extent to which statistical measures of some kind are used with respect to the entire logged output for the journal for each year.

The actual numbers of articles using any of the three statistical methods are also shown in white in the graph and positioned at the bottom of the respective bars for each year. This shows that the absolute numbers we are talking about here are often very small; we thus avoid perhaps artificially inflating their apparent contribution by scaling their internal proportion dimensions to 100%. In addition, since the bars show proportions with respect to the total number of logged articles for each year (which varies), their heights do not correspond directly to absolute counts either. For example, if we examine the two leftmost bars for *Visual Communication* (middle graph), we see first a bar showing exclusive use of ‘counting’ and then a bar showing exclusive use of ‘descriptive’ measures. These are the same height, indicating that they constitute the same proportions of the total logged output of that journal for those respective years (around 26%), but they correspond to *different* absolute numbers (2 and 1 respectively as shown in the graph) — from this we can see that the total number of logged articles for that journal for the first year was half of the total for the next year; this can also be read directly from the row of counts across the top of the graphs (6 and 3 respectively) as well as from the middle graph of Figure 1 above.

Taking all of the results together we can see, perhaps as would be expected, that the earlier articles all tended to offer either no numeric information or basic counts concerning the data. As time goes on there has been an overall increase in the use of inferential statistics among the papers identifying as empirical, particularly for *Visual Communication* and *Multimodal Communication*. *Social Semiotics* remains the journal where the least use is made of any statistical reporting beyond counts. As the figures in the graphs show, we are dealing with rather small numbers of absolute cases throughout and so any conclusions must be treated with caution. Nevertheless, we do seem to see a general, slow increase across the past two decades in the kinds and scales of empirical work being reported on in these journals. We take this as moderate support for our initial contention that the field of multimodality is, indeed, becoming more open towards empirically-relevant work and so it is, as a consequence, certainly worthwhile now considering in more detail just how that move can be best supported without losing contact with work that is not so inclined.

4 Promoting an Empirical Turn in Multimodality

As we have now shown, multimodality research, as represented in key journals of the field, has generally seen a steady increase in empirical work, including a more recent strengthening of the larger-scale quantitative line of research. Moreover, as we noted above, areas such as multimodal conversation analysis (Deppermann 2013), interaction studies (Mondada 2007, 2016), and others (see Section 3) have in any case pursued more quantitative approaches from early on. Nevertheless, as our study suggests, the field still shows a preference for qualitative, that is *explorative*, *descriptive* and *interpretative*, work on comparably small data sets.

As argued above (see also Bateman 2016: 37), such work is not, of course, any less revealing *per se* because of its qualitative nature. However, if left unaddressed, conceptual vagueness that may result from integrating frameworks from the various 'corners' of this highly interdisciplinary field may certainly restrict the explanatory 'reach' of such contributions. Indeed, there seems to be a general "lack of appropriate methodological guidance" in the field (Bateman 2016: 37, also in reference to Halliday 1994 and Forceville 2007); explicit discussions of *how precisely* to move '*from* theory to data' are rare (Bateman 2016: 37). Referring to cultural studies, Piper (2016: 6) describes this situation as "The Theory Gap". Furthermore, although ‘small n’ qualitative research is not a weakness of a disciplines’ empirical dimension in itself, it can become problematic if judgement continues to be led by untested intuition alone, particularly if we give in to the “temptation to generalize, to scale-up the nature of one’s argument” (Piper 2016: 4) even when those generalizations are not founded on a sufficient number of data points; Piper (2016: 4) calls this “The Evidence Gap”. Finally, there seems to be only a weak tradition of “documenting and theorizing our practices more extensively” (Piper 2016: 8, glossed as ‘The Self-Reflexive Gap’), especially when it comes to disclosing the details of methodological procedures of data collection and analysis.

It has to be emphasized, of course,that pursuing empirical approaches is “not a simple recipe for an unrealistically ‘clean’ structure of knowledge” (Sampson 2002: 5). In fact, empirical research usually leads to “a more puzzling picture” (van Peer et al. 2007: 21) since it makes us aware of the complexities of real-world data, or “[s]ometimes, nothing happens” (Piper 2016: 6), a negative result that may cause frustration. Both of these phenomena must be considered as gains rather than losses, because ‘a more puzzling picture’ may be precisely what we need to get us closer to the communicative reality we seek to understand; even ‘negative results’ are revealing and should be considered “as important as the novel insight of something previously unseen” (Piper 2016: 6). If we begin to close the Theory Gap, we are likely to produce more objective and reproducible descriptions of individual materials; if we start to close the Evidence Gap, we will be able to generalize our descriptions more reliably; and if we can close the Self-Reflexive Gap, we begin to mark out more clearly the “terrain of what one knows [...]” (Piper 2016: 8).

4.1 Quality Criteria in Other Empirical Fields and Their Transferability

Research communities with a more pronounced emphasis on empirical work have naturally engaged in discussions of how the quality of their research can be critically assessed. In psychology, for instance, it is common practice to evaluate the quality of empirical, often experimental, research designs on the basis of several criteria. Three primary criteria are: *objectivity*, *reliability*, and *validity* (see, e.g., Rost 2004; Himme 2009; Moosbrugger & Kelava 2014). There are also several secondary criteria: for instance, *scalability*, *test economy*, *practicability*, and *fairness* (see Himme 2009; Moosbrugger & Kelava 2014). The primary criteria, in particular, have important points of reference for other research fields with an interest in human behaviour, e.g., empirical social science research (see Kromrey 2002: 390–392), and so will also become relevant for multimodality research.

Objectivity is achieved if the test procedure (including the materials, the actual testing, and the generation and interpretation of results) is independent of any influences other than participant-specific factors, that is, independent of the researcher conducting the test, and the place and time at which the test is carried out (Rost 2004; Himme 2009; Moosbrugger & Kelava 2014; see also Krippendorff 2004). Objectivity can generally be imposed through standardization, e.g., by using a test manual with detailed instructions (Moosbrugger & Kelava 2014: 10). **Reliability** is smaller in scope, and zooms in on a test's measuring instruments and their capacity to produce the same results again and again upon repeating the test – independently of what the test is supposed to measure. Reliability is achieved when a test procedure produces highly correlating values across those variables that are assumed not to influence the test results and is typically assessed through retesting or parallel tests (Rost 2004; Himme 2009; Moosbrugger & Kelava 2014; see also Krippendorff 2004). Finally, **validity** concerns the content-related fit between what a test measures and what it is *supposed* to measure in light of the research questions. In that sense, it allows for estimating the meaningfulness of test results (Rost 2004; Himme 2009; Moosbrugger & Kelava 2014; see also Krippendorff 2004).

These three criteria are logically related to one another (Rost 2004: 33). Objectivity is a prerequisite for *reliability* because an accurate measuring instrument is useless if the reliable results it produced are not evaluated objectively (Rost 2004: 39). Also, objectivity and reliability may generally allow for a high accuracy of the measurements but the results are meaningless if they actually do not 'respond' to the research questions posed (Moosbrugger & Kelava 2014: 13).

Qualitatively-oriented research fields, to which many corners of current multimodality research are arguably closer, have engaged in extensive discussions of the transferability of these criteria for their own concerns. As a result, scholars have moved in different directions: those pursuing an *extrinsic* approach support importing criteria from quantitative research paradigms, and those pursuing an *intrinsic* approach suggest designing criteria exclusively for the qualitative context in which they are put to use (Sousa 2014: 213). It is important to note here that no approach is *per se* better than the other. After all, the quality of research needs to be assessed in relation to the research field framing it, its core interests, paradigms, and epistemology (cf. Sousa 2014: 212). Such a view grants qualitative fields that value an "on-site flexibility and less stepwise research design" (Yanow & Schwartz-Shea 2015: xix) the freedom to conduct case studies and practice "contexualized ('thick') description" (Bhattacharya 2008).

Multimodality research, in contrast, has always shown a strong interest in finding *regularities* and *patterns* in communicative behavior, not least due to its strong ties to linguistics, semiotics, and communication studies. An urge to theorize, to construct typologies and taxonomies, often leads multimodalists to abstract away from the singular and particular — and thus requires a corresponding approach for undertaking empirical research to support this. Such an approach would, ideally, on the one hand, aim to integrate qualitative perspectives productively. Such perspectives are still particularly relevant to empirical multimodality research at its current stage of development. They are “invaluable *tools for thinking semiotically* and can support useful conjectures, new conceptual arrangements, and are always ready to address new phenomena” (Bateman 2019: 315). On the other hand, due to our clear interest in finding productive generalizations, doing empirical multimodality research needs also to rely on quantitative methods to a greater extent, while paying more attention to making the move from theory to data (and back) explicit.

4.2 Five Criteria for Good Empirical Multimodality Research

The following five criteria are intended to describe quite explicitly what exactly makes certain research *empirical* in our understanding. At the same time, we hope thereby to provide some guidance as to what to consider when designing and conducting empirical research in multimodality. Similar to the quality criteria established in other fields, the five main criteria we foreground are:

1. *The Feedback Loop: From Theory to Data and Back*
2. *Objectivity*
3. *Reliability*
4. *Validity*
5. *Tentativeness of Results*

In addition, we also extend this list with several indicators for good empirical practice: *explicitness*, *transparency* (see, e.g., the open science movement), *replicability* & *replication*, *generalizability*, and *triangulation*. As can be seen below, we have positioned these criteria in relation to the main ones. For reasons of space, further criteria, such as *fairness*, *practicability*, or *sustainability*, are not addressed here, but are equally important and demand consideration whenever empirical research projects are being designed. At the end of this section, we then also give a short on note on the *data* needed for empirical analysis.

4.2.1 The Feedback-Loop: From Theory to Data and Back

From Theory to Data. The first half of the loop requires a detailed operationalization of broader theoretical constructs in order to make them “reliably recoverable”

(Bateman 2019: 303) and thereby to ensure descriptions of those phenomena they are supposed to describe. Within *Legitimation Code Theory* (Maton & Chen 2016), several useful mechanisms have been described for these purposes: *Data instruments* provide methodological recommendations as to how abstract theoretical concepts suggest “foci for data collection and questions for analysis” (Maton & Chen 2016: 30). *Mediating languages* constitute typologies of categories that serve to make theoretical concepts more sensitive to the particularities of actual data. And *translation devices* operate at a low level of abstraction and are sensitive to the context of a particular study (Maton & Chen 2016: 31). Employing these mechanisms, and being explicit about how they were employed, ensures that the connection between theory and data can be made in a *reliable* fashion, and their data-sensitiveness helps enforcing preconceived theoretical concepts (see Bateman 2019: 301).

From Data to Theory. The second half of the loop is concerned with processing the annotated data and relating the results back to theory. This can be done efficiently by a search for patterns and regularities, typically accomplished in a top-down (theory-driven) or bottom-up (data-driven) fashion (see Bateman & Hiippala, this volume). Such methods involve quantification and an exploration of *correlations* across various kinds of descriptions (Bateman 2014: 252). Statistical processing is not at all limited to numerical descriptions; all it takes to involve category-based annotations is a statistical model fit to process them (see also Bateman & Hiippala, this volume).

Statistical approaches typically require larger quantities of data to produce meaningful results. Thus, statements about the general validity of smaller-scale studies need to be made with appropriate caution. Even if annotations produced on the basis of a single text suggest a mismatch between previous theory and features of actual data, this indication remains weak until backed up with further results. Thus, corpus work needs to become larger in scale, which, in turn, means relying more than before on (semi-)automated analyses and visualization methods (cf. O’Halloran et al. 2011; Bateman 2014: 252; Kohrs 2018). Our community needs to continue working towards making such methods more accessible to fellow scholars, while promoting the pooling of our various skill sets in broader research teams.

4.2.2 Objectivity

Our investigations of multimodal artifacts or performances may be classified as *objective* to the extent that our frameworks can be applied without any more particular knowledge beyond what is specified in a test manual. Ideally, then, any analyst can apply the framework, at any place, or time with similar outcomes. Objectivity can thus be achieved if we are sufficiently *explicit* about previous assumptions or knowledge necessary for data collection and analysis (*explicitness*). One technique commonly used for this, particularly in content analysis (cf. Schreier 2012), is to produce a so-called *code book* where one can document previous assumptions, methodological recommendations (*data instruments*), and typologies of analytical categories (*mediating languages* and *translation devices*).

It is increasingly seen as good practice in many fields to be *transparent* by making such code books *publicly* available, together with the actual data, all documentation regarding the research questions/hypotheses, the choice of methods, the annotation process, and even further processing steps (including code for statistics software such as *R*). All such considerations become ever more important when engaging with interdisciplinary work (see Yanow & Schwartz-Shea 2015: xv; see also Sousa 2014: 216).

If justice is done to the quality criterion of *objectivity*, a multimodal study not only becomes *repeatable* by other researchers (*replicability*), it also allows them to challenge previous research designs, to correct them, or to build on them (see Piper 2016: 7 on “The Self-Reflexivity Gap”).

4.2.3 Reliability

Much of contemporary non-experimental multimodality research accomplishes the description of data through applying conceptual categories. In this context, *reliability* refers to whether a concept and its associated definition is capable of producing the same categorizations repeatedly when used to describe the same phenomena in similar data sets.[2] A high degree of reliability can be achieved by working out in detail how theoretical concepts are to be operationalized in their application to data, for example in the form of a code book as mentioned above. Training sessions in which coders annotate smaller data sets can be used to assess the reliability of an annotation scheme through *inter-coder reliability* checks (Krippendorff 2004: 215). Also, researchers can test if the same results are generated upon repeating a study (testing for *intertemporal stability*). This also provides an opportunity to revisit decisions made in designing and conducting a study.

[2] Evidently, analytical work in multimodality extends beyond the application of coding schemes. Thus, analytical tools are not only categorical but include, e.g., rulers used to measure layout spaces, color pickers used to determine saturation measures, or eye-trackers used to record gaze saccades. Since the use of such measuring tools can typically be made highly reliable, we do not discuss them further at this point.

Benoit & Holbert (2008: 615–616) argue that, while repeating a study is common practice in the natural or empirical social sciences, *replication* still has to gain traction in communication studies and other humanities(-related) research fields. Making research in multimodality more reliable requires explicit documentation (*objectivity*), even if this entails taking up 'valuable journal space' (Benoit & Holbert 2008: 616). And even prior to publishing a study's results, this can require considerable resources to conduct tests of the intersubjective stability of any coding used.

This may well be beyond the scope of individual research projects, particularly at graduate level, and so our recommendations here are straightforward. Even if the limited size of a research project makes it difficult to pursue 'double-coding', this goal should nevertheless be borne in mind as an 'ideal' that one is, for perfectly justifiable reasons, perhaps not achieving in some particular case. If one designs a project as far as possible so that double-coding *could* have been done, then the resulting design will be more likely to satisfy the other criteria more fully as well. We should generally re-think multimodality research as a 'team effort' even when the team remains unrealized. In any case, one should report whether reliability tests were conducted and, if not, indicate what made that difficult or impossible; sometimes research might simply be too exploratory to warrant reliability checking. Explicitness concerning this point is always preferable.

4.2.4 Validity

As in empirical testing, achieving *validity* in an empirical multimodal study hinges on how well our research project scores with respect to the criteria of *objectivity* and *reliability*. Even if the descriptive categories correspond to the research question, the findings are unusable if they turn out to be too vague to produce *reliable* categorizations. Likewise, if a coder lacks a solid grasp of the coding scheme (even if it would have afforded *reliability*), and their annotations are thus skewed by a subjective 'interpretation' of the concepts, the study will not afford valid results. Paying close attention to carrying out a research project *objectively* and using *reliable* tools therefore forms a solid basis for achieving *validity*.

Validity can be ensured if researchers are explicit (*explicitness*, *objectivity*) about their research questions and hypotheses, while making ample reference to existing knowledge in whatever area of investigation is at issue (*construct validity*, Moosbrugger & Kelava 2014: 16, see also Krippendorff 2004: 315). On this basis, researchers should then seek to compile annotation schemes in such a way that the phenomena under investigation are *adequately represented*. An estimation of *adequacy*, in this context, is difficult to measure and thus derived from an 'informed judgement' based on the knowledge shared within a research community (*content validity*, Moosbrugger & Kelava 2014: 15, see also Krippendorff 2004: 315).

4.2.5 Tentativeness of Results

The quality criterion seems simple since it is a presupposition of the idea of working scientifically. Upon closer scrutiny, however, both the idea and its consequences are far from trivial: the results we generate through empirical (multimodality) research are always *tentative* in nature, and may need to be replaced at some point (see Sampson 2005: 4; Sousa 2014: 217). This ties in with Peirce's notion of *pragmatism* and is also a necessary consequence of his concept of *abduction*, the process of forming explanatory hypotheses (Peirce 1931–1958: 5.172). The tentativeness of results is at the very heart of empirical research — a constant invitation to travel along the path of the feedback loop, and be more 'knowledgeable' every time one comes round full circle.

4.2.6 A Final Note on Data

In addition to being intersubjectively accessible, the *data* for multimodal studies should be selected in view of the research question or hypotheses so that it affords providing answers to the questions posed (see Bateman & Hiippala, this volume). Once the type of data has been decided on, it is usually collected into a *corpus* (Bateman 2014: 239)[3]; many examples are given in the chapters of this book.

When conducting empirical research projects, researchers face the (vexed) question of how much data they will need in order to make any valid claims. The amount of data that one needs to find an effect depends on how frequently the effect

[3] A *corpus* is a collection of data specifically selected and further structured in view of a given research question or hypothesis, whereas *data* is a more general expression typically used to denote any kinds of materials that are studied as part of empirical work. Data is also sometimes used to describe a selection of measurements/values gained through the empirical study of a given corpus or obtained in experimental studies. The further processing of corpora is usually done with the help of *corpus tools*. While there is a plethora of tools available for annotating and analysing *linguistic* corpora, multimodality research faces difficulties with many fewer and only rather specific tools available (Bateman 2013). These need to cater for multi-level annotations that provide information concerning various facets of the materials under study, ranging from technical features, to transcriptions of selected perspectives on the data, to results of experimental studies, to hypotheses about category attributions to individual segments or units, and the relationships between them (Bateman 2014: 251).

occurs: if it occurs very often, then obviously it is more likely that some collection of data will include sufficient examples to draw conclusions. The amount of data required also depends on just how ‘strong’ an effect is – if it is a strong effect, then, again, less data will be necessary to show it at work. In any case, the data gathered needs to include at least data exhibiting the range of variations and phenomena that are the target of the research questions. For example, if one is probing the different use of multimodal resources made by contrasting groups of sign-users, then the data needs to include sufficient examples from those groups, and so on. There is no point probing data for variation that the data does not include.

5 Overview of Contributions

The particular view and methodological requirements of *empirical multimodality research* that we have laid out in this introduction are meant to provide a more general thematic frame in which to place the typically much more specific discussions and case studies presented in the remainder of this book. Following on the present chapter, which makes up *Part I — Introduction*, the book continues with two further parts: *Part II — Charting Paths for Empirical Research: Theoretical and Methodological Reflections* (Chapters 1–4) and *Part III — Empirical Inroads: Case Studies and Results* (Chapters 5–10). Part II provides insights in theoretical thoughts and methodological discussions of recent contributions to the field of empirical multimodality research; Part III provides some rich empirical case studies of multimodal artefacts and performances to illustrate state-of-the-art empirical work in the field and to identify persisting research gaps and suggest future avenues for research. In this section, we survey the contributions briefly and position them in relation to the argumentation presented above.

Part II — Charting Paths for Empirical Research: Theoretical and Methodological Reflections.

In his chapter entitled *Dimensions of Materiality*, **John A. Bateman**’s starting point is the long-standing tradition of attributing *materiality* a central role in multimodality research. The development of an empirically robust account of materiality is then central. To this end, he argues that ‘external languages of description’ (cf. Maton & Chen 2016) are needed for securing and organizing proper analytical ‘access’ to data. On the basis of previous work in Bateman et al. (2017), the chapter construes materiality as such as an external language of description, and introduces temporality, space, role, and transcience as its central characteristics. This extended view of materiality is finally related to the semiotic purposes of communication in the broad framework of multimodality adopted and illustrated by the example of three different communicative situations and their comparison with regard to their material canvases. Thus, the chapter provides a systematic approach to materiality as a "reliably recoverable" (Bateman 2019: 303; see above) theoretical construct. The author's contribution is thus an important step toward robust empirical methodologies and to achieving a close connection between theory and data.

Pursuing similar aims, **John A. Bateman** and **Tuomo Hiippala**'s chapter, *From Data to Patterns*, sheds light on the concept and practice of *modeling* in empirical research. With the aim of crossing the disciplinary boundaries in multimodality research, the authors suggest an understanding of models as specifically structured descriptions of patterns and regularities from a semiotically oriented perspective on iconicity, following Peirce. On this theoretical basis, the empirical procedure of moving from theory to models to data and back is further discussed and exemplified by a critical evaluation of certain types of modeling procedures and techniques for formulating and evaluating models.

The following two chapters then combine theoretical discussions with more extensive practical work. **Barbara Tversky** and **Angela Kessell**'s chapter, entitled *Thinking in Action*, focuses on the mapping of thoughts to non-verbal and verbal expressive resources. Based on their empirical work, the authors demonstrate that gestures and marks on a sheet of paper have many important properties in common. They furthermore bring out that gesture supports direct (iconic) expression of actions. This engagement of the embodied perception of action as well as of the visual offers considerable benefits in relation to language alone. The authors consequently argue for a combined network of gesture, action, the designed world, and abstraction, which they call 'spraction'. Despite being a reprint of a previously published journal article (2014), the contribution offers invaluable insights which are as topical today as then, particularly in the context of current multimodality discussions.

Finally in Part II, **Ralph Ewerth**, **Christian Otto**, and **Eric Müller-Budack**'s chapter, *Computational Approaches for the Interpretation of Image-Text Relations*, discusses and demonstrates computational approaches to the processing and interpretation of text-image relations from a computer science perspective. Based on previous work on the classification of text-image relations, and taking into account approaches from linguistics and communication studies as well, the authors define computable dimensions for different types of relations, namely cross-modal mutual information, semantic correlation, and status relation, and use these to draw out eight image-text classes and their relation to existing taxonomies. The chapter furthermore reports on experimental results generated through an automatic classification of these text-image relations applying deep learning approaches and presents a more differentiated model for the dimension of cross-modal information for image-text pairs in news. The approaches developed contribute to the expansion of tried-and-tested empirical methodologies for multimodality research that utilize automated computer-based processing to bridge the 'semantic gap' between text and image.

Part III — Empirical Inroads: Case Studies and Results.

In his chapter *"I can't see why you're laughing": Multimodal Analysis of Emotionalized Political Debate*, **Andreas Rothenhöfer** places an extract from a UK TV news programme showing a short interaction between a politician and a news anchor under the analytic microscope. This extract received considerable public attention due to the politician's supposedly controversial *smirk*-like reaction. To understand the interaction and its take-up in more detail, Rothenhöfer pursues a mixed-methods approach to facial expression analysis in combination with a qualitative pragmatic perspective with the aim of analysing the reception, co-construction, and reframing of the short interview sequence as presented on Twitter. By using the computational platform *iMOTIONS* and the analytical software *Affectiva*, the author demonstrates the usefulness and applicability of biometric analysis to reconstruct and distinguish behavioral interaction chains from more general mood or attitude aspects, and to support or contradict the perception of such interactions in public discourse. The study thus also contributes to a further exploration of software-based tools and their productive complementation with qualitative approaches.

In their chapter entitled *A Corpus-based Approach to Color, Shape, and Typography in Logos*, **Christian Mosbæk Johannessen**, **Mads Lomholt Tvede**, **Kristoffer Claussen Boesen**, and **Tuomo Hiippala** then present a data-driven corpus study of color, shape, and typography in corporate logos. With the aim of addressing the 'social style' of logos as representations of brands and branding, the authors operationalize their analysis of the graphic canvases of 50 logos from the oil industry and non-governmental environmental organizations by analyzing the dimensions of shape, color, and typography. The analytical framework reflects 14 types of material properties, and positions them as variables whose values represent stylistic choices in logo design. For an evaluation of the interaction of these variables, the authors employ Principal Component Analysis (PCA) as an inferential statistical method to exhibit patterns of variation in the corpus. Results show that certain groups of logos show statistically significant differences in their specific uses of shapes, color, for example, and suggest that that variation is dependent on the organization/industrial sector they are used to represent.

In the following contribution, entitled *Pixel Surgery and the Doctored Image*, **Hartmut Stöckl** draws on a corpus of 232 print advertisements from *Lürzer's Archive* in order to investigate the function of computer-generated images when used to construct multimodal arguments. Combining scholarship in pictorial theory, visual rhetoric, and multimodal argumentation, the author develops an extensive typology of manipulations of visual structure ('design operations'), investigates the rhetorical potential they bear, their relational propositions and, ultimately, the argument types they support. Stöckl's contribution features a detailed code book which noticeably increases the study's degree of objectivity. The relative frequencies of occurrences of the annotated categories are then interpreted as prototypical and functionally effective image design and multimodal argumentation strategies pursued in advertising.

Next, **Jiaping Kang** and **Zhanhao Jiang**'s chapter, entitled *Multimodal Discourse Analysis Based on the GeM Model*, presents a thorough application of the *Genre and Multimodality*-framework (originally designed for analyzing page-based documents: cf. Bateman 2008) to a small corpus of 10 U.S.-American and Chinese environmental protection posters. With the help of XML-coding to annotate the posters' basic compositional unity, their layout and rhetorical structure, and utilizing the GeM-Tools developed for further processing by Hiippala (2015), the authors provide a contrastive analysis of the semiotic resources used in both sets of posters. The results show, for example, that the distribution of verbal and visual units in the two cultures is very similar, but that there is variation in the use of language and typography. Despite its comparably small sample size and leaning towards qualitative empirical research, the study clearly illustrates the gains of a detailed multi-layer analysis that seeks to tie lower-level data-sensitive annotations to higher-level analytical concepts, such as rhetorical relations, and so stands well as a motivation for potentially larger-scale studies.

Adopting a quite different theoretical approach, **Loli Kim** and **Jieun Kiaer** draw in their chapter, *Conventions in How Korean Films Mean*, on the framework of Segmented Film Discourse Representation Structures (Wildfeuer 2014) to conduct a pilot study of the nature and content of 'final goodbye'-events in the contemporary South Korean films *Old Boy* (2003), *Sympathy for Lady Vengeance* (2005), and *The Man from Nowhere* (2010). By formally specifying discourse segments and discourse relations in several relevant film scenes, the authors identify reoccurring patterns of filmic configurations that can be assumed to function as conventions among the three films analyzed. The results show that such empirically-supported testing of existing methodological frameworks adds considerable detail and precision to the understanding of how meaning in film is constructed.

Finally, **Dušan Stamenković** and **Janina Wildfeuer**'s contribution, entitled *An Empirical Multimodal Approach to Open-World Video Games*, reports on a case study analysis of the video game *Grand Theft Auto V* (Rockstar North 2013). The authors present their extensive annotation work with regard to all 80 main missions of the game, and draw out a comprehensive semiotic inventory of the game's basic semiotic elements used in these missions. These elements are further analysed with regard to their frequency of occurrence in the game to show statistically attestable associations between variables (correlations). On this basis, Stamenković and Wildfeuer demonstrate how a diversity of combined features structure the experience of playing the game, and guide and instruct players within an essentially open game world. They also show how sufficient empirical evidence for specific semiotic elements in complex multimodal artefacts establishes a more stable ground for investigating hypotheses of meaning-making in these artefacts.

6 Framing Conclusions

Despite the natural interest of empirical multimodality research in conducting data-based research, the approach is certainly still far from realizing its full potential. In this introduction, we have discussed the gradual implementation of an increasing range of quantitative work as well as larger-scale studies and have argued that this now constitutes an important avenue to pursue for the advancement of the field. At the same time, however, this must be done while still granting more exploratory work a permanent and prominent position in the overarching research agenda. Given the breadth of multimodality concerns, there will always be a need for exploration: what we suggest, however, is that even exploration can be undertaken with an eye to subsequent, less exploratory investigations in depth.

Consequently, we also argued further that, in order to ultimately achieve a more robust empirical grounding for multimodality research across the board, both qualitatively- and quantitatively-oriented studies would benefit from allowing themselves to be guided by five core quality criteria for good empirical practice — namely completing the feedback-loop (from theory to data and back), implementing the principles of objectivity, reliability, and validity, and acknowledging that the results generated will necessarily remain tentative in nature. We believe that following these principles in future research is essential if we are to continue our productive investigations of increasingly complex artefacts and performances, to further strengthen our theoretical and methodological frameworks, and to ultimately encourage an even more intense exchange among the diverse communities within our emerging discipline, and beyond. Particular examples and directions for these developments are evident in all of the individual contributions to the volume.

Acknowledgements
We wish to thank Hartmut Stöckl (Salzburg) and Bradley K. Mackay (Salzburg) for their insightful comments on an earlier version of this manuscript.

Bibliography

Baldry, Anthony & P. J. Thibault. 2005. Multimodal Corpus Linguistics. In G. Thompson & S. Hunston (eds.), *System and Corpus: Exploring Connections*, 164–183. London and New York: Equinox.

Baldry, Anthony & P. J. Thibault. 2006. *Multimodal Transcription and Text Analysis: A Multimedia Toolkit and Coursebook with Associated On-line Course*. Textbooks and Surveys in Linguistics. London and New York: Equinox.

Bateman, John A. 2008. *Multimodality and Genre: A Foundation for the Systematic Analysis of Multimodal Documents*. Basingstoke: Palgrave Macmillan.

Bateman, John A. 2013. Multimodal Corpus-Based Approaches. In C. A. Chapelle (ed.), *The Encyclopedia of Applied Linguistics*, Hobeken, NJ, USA: Blackwell Publishing Ltd. https://doi.org/10.1002/9781405198431.wbeal0812.

Bateman, John A. 2014. Using Multimodal Corpora for Empirical Research. In C. Jewitt (ed.), *The Routledge Handbook of Multimodal Analysis*, 238–252. London: Routledge 2nd edn.

Bateman, John A. 2016. Methodological and Theoretical Issues for the Empirical Investigation of Multimodality. In N.-M. Klug & H. Stöckl (eds.), *Handbuch Sprache im multimodalen Kontext* (Handbooks of Linguistics and Communication Science (HSK) 7), 36–74. Berlin: De Gruyter Mouton.

Bateman, John A. 2019. Afterword: Legitimating Multimodality. In J. Wildfeuer, J. Pflaeging, J. Bateman, O. Seizov & C. Tseng (eds.), *Multimodality: Disciplinary Thoughts and the Challenge of Diversity*, 297–321. Berlin: De Gruyter Mouton.

Bateman, John A., J. L. Delin & R. Henschel. 2004. Multimodality and Empiricism: Preparing for a Corpus-Based Approach to the Study of Multimodal Meaning-Making. In E. Ventola, C. Charles & M. Kaltenbacher (eds.), *Perspectives on Multimodality*, 65–87. Amsterdam: John Benjamins.

Bateman, John A., J. Wildfeuer & T. Hiippala. 2017. *Multimodality – Foundations, Research and Analysis. A Problem-Oriented Introduction*. Berlin: De Gruyter Mouton.

Bednarek, Monika. 2015. Corpus-Assisted Multimodal Discourse Analysis of Television and Film Narratives. In P. Baker & T. McEnery (eds.), *Corpora and Discourse Studies: Integrating Discourse and Corpora*, 63–87. London: Palgrave Macmillan. https://doi.org/10.1057/9781137431738_4.

Benoit, William L. & R. L. Holbert. 2008. Empirical Intersections in Communication Research: Replication, Multiple Quantitative Methods, and Bridging the Quantitative-Qualitative Divide. *Journal of Communication* 58(4). 615–628.

Bezemer, Jeff & K. Cowan. 2021. Exploring Reading in Social Semiotics: Theory and Methods. *Education 3-13* 49(1). 107–118. https://doi.org/10.1080/03004279.2020.1824706.

Bhattacharya, Himika. 2008. Empirical Research. In *The SAGE Encyclopedia of Qualitative Research Methods*, 253–255. Los Angeles: Sage.

Bold, Richard A. & E. Herranz. 1992. Two-Handed Gesture in Multi-Modal Natural Dialog. In *ACM UISZ '92 Symposium on User Interface Software and Technology*, 7–14. ACM Press.

Bollnow, Otto Friedrich. 1974. The Objectivity of the Humanities and the Essence of Truth. *Philosophy Today* 18(4). 3–18.

Brannen, Julia. 1992. *Mixing Methods: Qualitative and Quantitative Research*. London: Avebury.

Bühler, Axel. 2003. Grundprobleme der Hermeneutik. In A. Bühler (ed.), *Hermeneutik: Basistexte zur Einführung in die wissenschaftstheoretischen Grundlagen von Verstehen und Interpretation*, 3–19. Heidelberg: Synchron.

Carter, Ronald & S. Adolphs. 2008. Linking the Verbal and the Visual: New Directions for Corpus Linguistics. *Language and Computers* 64. 275–291.

Deppermann, Arnulf. 2013. Multimodal Interaction from a Conversation Analytic Perspective. *Journal of Pragmatics* 46(1). 1–7.

Engels, Tim CE, A. I. Starčič, E. Kulczycki, J. Pölönen & G. Sivertsen. 2018. Are Book Publications Disappearing from Scholarly Communication in the Social Sciences and Humanities? *Aslib Journal of Information Management* 70(6). 592–607.

Forceville, Charles J. 2007. Book Review: *Multimodal Transcription and Text Analysis: A Multimedia Toolkit and Coursebook* by Anthony Baldry and Paul J. Thibault. *Journal of Pragmatics* 39(6). 1235–1238.

Giard, M. H. & F. Peronnet. 1999. Auditory-Visual Integration during Multimodal Object Recognition in Humans: A Behavioral and Electrophysiological Study. *Journal of Cognitive Neuroscience* 11(5). 473–490. https://doi.org/10.1162/089892999563544.

Goodwin, Charles. 1980. Restarts, Pauses, and the Achievement of Mutual Gaze at Turn-Beginning. *Social Inquiry* 50. 272–302.

Gu, Yueguo. 2006. Multimodal Text Analysis: A Corpus Linguistic Approach to Situated Discourse. *Text & Talk* 26(2). 127–167.

Halliday, Michael A. K. 1994. Systemic Theory. In R. Asher (ed.), *The Encyclopedia of Language and Linguistics*. Oxford: Pergamon Press.

Hiippala, Tuomo. 2015. *The Structure of Multimodal Documents: An Empirical Approach*. London: Routledge.

Himme, Alexander. 2009. Gütekriterien der Messung: Reliabilität, Validität und Generalisiertbarkeit. In S. Albers, D. Klapper, U. Konradt, A. Walter & J. Wolf (eds.), *Methodik der Empirischen Forschung*, 485–500. Wiesbaden: Springer.

Jewitt, Carey. 2017. Multimodal Discourses across the Curriculum. In S. L. Thorne & S. May (eds.), *Language, Education and Technology* (vol. 9 Encyclopedia of Language and Education), 31–43. Cham: Springer 3rd edn.

Kevles. 1992. Historical Foreword. In H. Robin (ed.), *The Scientific Image: From Cave to Computer*, 10–19. New York: Freeman and Abrams.

Kirk, Jerome & M. L. Miller. 1986. *Reliability and Validity in Qualitative Research*. London: Sage.

Kohrs, Kirsten. 2018. Learning from Linguistics: Rethinking Multimodal Enquiry? *International Journal of Social Research Methodology* 21(1). 49–61.

Koons, D.B., C. Sparrel & K. Thorisson. 1993. Integrating Simultaneous Input from Speech, Gaze, and Hand Gestures. In M. Maybury (ed.), *Intelligent Multimedia Interfaces*, 257–276. MIT Press.

Krippendorff, Klaus. 2004. *Content Analysis: An Introduction to its Methodology*. London and Thousand Oaks, CA: Sage 2nd edn.

Kromrey, Helmut. 2002. *Empirische Sozialforschung*. Opladen: Verlag Leske + Budrich.

Maton, Karl & R. T.-H. Chen. 2016. LCT in Qualitative Research: Creating a Translation Device for Studying Constructivist Pedagogy. In K. Maton, S. Hood & S. Shay (eds.), *Knowledge-Building: Educational studies in Legitimation Code Theory*, 27–48. Abingdon and New York: Routledge.

Mondada, Lorenza. 2007. Multimodal Resources for Turn-taking: Pointing and the Emergence of Possible Next Speakers. *Discourse Studies* 9(2). 195–226.

Mondada, Lorenza. 2016. Challenges of Multimodality: Language and the Body in Social Interaction. *Journal of Sociolinguistics* 20(2). 2–32.

Moosbrugger, Helfried & A. Kelava. 2014. Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger & A. Kelava (eds.), *Testtheorie und Fragebogenkonstruktion*, 8–26. Heidelberg: Springer.

Nakano, Yukiko & M. Rehm. 2009. Multimodal Corpus Analysis as a Method for Ensuring Cultural Usability of Embodied Conversational Agents. In M. Kurosu (ed.), *Human Centered Design. International Conference on Human Centred Design*, 521–530. Berlin and Heidelberg: Springer.

Norris, Sigrid. 2002. The Implication of Visual Research for Discourse Analysis: Transcription Beyond Language. *Visual Communication* 1(1). 97–121.

O'Halloran, Kay L., S. Tan, B. Smith & A. Podlasov. 2011. Challenges in Designing Digital Interfaces for the Study of Multimodal Phenomena. *Information Design Journal* 18(1). 2–21.

Pederson, Kaitlin & N. Cohn. 2016. The Changing Pages of Comics: Page Layouts across Eight Decades of American Superhero Comics. *Studies in Comics* 7(1). 7–28.

van Peer, Willie, F. Hakemulder & S. Zyngier. 2007. *Muses and Measures: Empirical Research Methods for the Humanities*. Cambridge: Cambridge Scholars Publishing.

Peirce, Charles Sanders. 1931–1958. *Collected Papers of Charles Sanders Peirce*. Cambridge, MA: Harvard University Press. Vols. 1–6, 1931-1935, edited by Charles Hartshorne and Paul Weiss; Vols. 7–8, 1958, edited by Arthur W. Burks.

Piper, Andrew. 2016. There Will Be Numbers. *Journal of Cultural Analytics* 1. 1–10.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/.

Riesenhuber, Felix. 2009. Großzahlige empirische Forschung. In S. Albers, D. Klapper, U. Konradt, A. Walter & J. Wolf (eds.), *Methodik der empirischen Forschung*, 1–16. Wiesbaden: Springer.

Rost, Jürgen. 2004. *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Verlag Hans Huber.

Sampson, Geoffrey. 2002. *Empirical Linguistics*. London: Continuum.

Sampson, Geoffrey. 2005. Quantifying the Shift towards Empirical Methods. *International Journal of Corpus Linguistics* 10(1). 15–36.

Schreier, Margrit. 2012. *Qualitative Content Analysis in Practice*. London: Sage.

Snow, Charles Percy. 2001 [1959]. *The Two Cultures*. London: Cambridge University Press.

Sousa, Daniel. 2014. Validation in Qualitative Research: General Aspects and Specificities of the Descriptive Phenomenological Method. *Qualitative Research in Psychology* 11. 211–227.

Stöckl, Hartmut. 1997. *Textstil und Semiotik englischsprachiger Anzeigenwerbung*. Frankfurt am Main: Peter Lang.

Streeck, Jürgen. 1983. Konversationsanalyse. Ein Reparaturversuch. *Zeitschrift für Sprachwissenschaft* 2(1). 72–104. https://doi.org10.1515/ZFSW.1983.2.1.72.

Thibault, Paul J. 2000. The Multimodal Transcription of a Television Advertisement: Theory and Practice. In A. P. Baldry (ed.), *Multimodality and Multimediality in the Distance Learning Age*, 311–385. Campobasso, Italy: Palladino Editore.

Thomas, Martin. 2019. Making a Virtue of Material Values: Tactical and Strategic Benefits for Scaling Multimodal Analysis. In J. Wildfeuer, J. Pflaeging, J. A. Bateman, O. Seizov & C.-I. Tseng (eds.), *Multimodality. Disciplinary Thoughts and the Challenge of Diversity*, 69–92. De Gruyter Mouton.

Thomson, T.J. 2021 forthcoming. International, Innovative, Multi-Modal, and Representative? The Geographies, Methods, Modes, and Aims Present in two Visual Communication Journals. *Visual Communication*.

Thorisson, Kristinn R., D. B. Koons & R. A. Bolt. 1992. Multi-Modal Natural Dialogue. In *Human Factors in Computing Systems, CHI 92 Conference Proceedings*, 653–654. ACM Press.

Wickham, Hadley. 2016. Programming with ggplot2. In *ggplot2: Elegant Graphics for Data Analysis*, 241–253. Springer. https://doi.org/10.1007/978-3-319-24277-4_12.

Wildfeuer, Janina. 2014. *Film Discourse Interpretation. Towards a New Paradigm for Multimodal Film Analysis*. London and New York: Routledge.

Wildfeuer, Janina, J. Pflaeging, J. Bateman, O. Seizov & C. Tseng. 2019. Multimodality: Disciplinary Thoughts and the Challenge of Diversity. Introduction. In J. Wildfeuer, J. Pflaeging, J. Bateman, O. Seizov & C. Tseng (eds.), *Multimodality: Disciplinary Thoughts and the Challenge of Diversity*, 3–38. Berlin: De Gruyter Mouton.

Yanow, Dvora & P. Schwartz-Shea. 2015. Wherefore 'Interpretative'? An Introduction. In D. Yanow & P. Schwartz-Shea (eds.), *Interpretation and Method: Empirical Research Methods and the Interpretative Turn*, xiii–xxxi. New York: Routledge.

Part II: **Charting Paths for Empirical Research: Theoretical and Methodological Reflections**

John A. Bateman

Dimensions of Materiality

Towards an External Language of Description for Empirical Multimodality Research

Abstract: The field of multimodality currently faces a double challenge: first, the datasets drawn on for bolstering argumentation need to grow in order to better support empirical investigation; and second, the communicative situations and artifacts addressed by the field are themselves becoming ever more complex. These demands raise a multitude of issues with methodological consequences for the everyday practice of analysing multimodal phenomena. In this chapter, it is argued that a thorough semiotic re-engagement with the nature of materiality and the distinct kinds of traces that materialities can support offers a powerful analytic technique for securing access to data regardless of how multimodally complex such data become. The chapter construes its account of materiality as an 'external language of description' in the sense of Legitimation Code Theory, by which analysis of multimodal data can proceed without presupposing the very theoretical categories for which empirical support is being sought. Several examples are discussed and the relevance of a more finely articulated notion of materiality for drawing connections between superficially quite different communicative situations demonstrated.

Keywords: materiality, canvas, external language of description, Legitimation Code Theory, semiotic traces, empirical methods

1 Introduction: The Material Turn

Over the past two decades most existing approaches to multimodality research have come to accord *materiality* a central role (cf. e.g., van Leeuwen 1999; Scollon 2001; Kress & van Leeuwen 2001; van Leeuwen 2009; Norris 2009; Iedema 2007; Björkvall & Karlsson 2011; Streeck 2013; Pirini 2016; Johnson 2018; Bateman 2019b). In this literature materiality is invoked not in the sense of physics, but rather as part of a general re-appraisal of the importance of embodiment and engagement with physical objects and their environments for almost all aspects of meaning-making. This 'material turn' is by no means limited to multimodality studies (cf. e.g., Gottdiener 1995; Mersch 2002; Hayles 2003; Latour 2005; Orlikowski 2007; Drucker 2013; Elleström 2014; Mukerji 2015) and, indeed, has an even longer history

https://doi.org/10.1515/9783110725001-002

in discussions of the materiality of communication and media. Proclamations of rich entanglements between the neurocognitive, the social, and the material are now commonplace.

The goal of this chapter will be to develop materiality further as an explicit component of a robust empirical methodology for multimodality studies. This will be pursued from two perspectives. First, focusing attention on materiality naturally brings into close relief those very 'objects of analysis' (construed quite literally) that are of central concern for multimodality. It will consequently be argued that a better understanding of materiality contributes directly to methodology in that knowing more about materiality also supports more robust and well designed empirical studies. Second, an appropriate account of materiality can also be productively theorized at a considerably deeper methodological level as an answer to a basic question facing any approach to empirical research: the question of how to *secure access to data at all*. These perspectives, although clearly related, address rather different concerns. Whereas the explicit account of materiality set out provides specific guidance for conducting empirical multimodality research, the issue of securing access to data contributes to a broader conceptualization of the entire nature of multimodality as a field of inquiry.

The organization of the chapter will lead us through these two contributing perspectives. First, the chapter will approach the general problem of securing access to data from the position of Karl Maton and colleagues' Legitimation Code Theory (LCT: Maton 2014, 2016). According to LCT, empirical disciplines must provide explicit accounts of how they go about securing access to data in order to ground their research practices effectively. To achieve this, disciplines need to define 'external languages of description' that mediate between theory and data. Second, the chapter will build on the proposal made in Bateman (2019a) that the classification of materiality set out by Bateman et al. (2017: 101–110) offers precisely such an external language of description for multimodality research. By these means, we will see how an appropriate articulation of material possibilities can be mobilized as a powerful tool for organizing data prior to further study. The chapter will then show this in operation with respect to several examples of multimodal communication involving more or less complex materialities. In each case, it will be argued that construing a characterization of materiality as an external language of description provides critical guidance for subsequent empirical investigation. A brief restatement then concludes the chapter.

2 External Languages of Description and Accessing Data

Arguably one of the most essential challenges facing the practices within any discipline is how those practices can be made to serve an enabling function for that discipline's community's *construction of knowledge over time*. To be effective, strategies for best achieving this construction must be made explicit and situations that might hinder this diagnosed and corrected. In all empirical work, one needs to be able to generalize across specific cases by applying analytic categories motivated by theory, but this is in itself a complex undertaking. In particular, it is important to avoid the often inter-related methodological problems raised by what Bateman et al. (2017: 231) term the 'description trap' and the 'pseudotechnicality trap'. Here analysis is replaced by detailed descriptions of data that simply employ the terms given by theory. This reduces analysis to a labelling exercise which may give an impression of technicality, but which can equally fail to reveal areas of phenomena where the categories of the theory do *not* fit. Empirical methodologies need then to be developed so that it is possible to see where theoretical constructs may actually be being challenged by data.

Legitimation Code Theory, an account of the mechanisms of disciplinary knowledge-building drawing on both the sociologist Basil Bernstein and the sociologist and philosopher Pierre Bourdieu, addresses many of these concerns explicitly. Two essential tasks are defined. First, appropriate organizational forms must be found for the knowledge itself. And second, effective ways must be found for relating this knowledge to its objects of concern, whatever those may be. Concerning the first task, Bernstein, and later Maton within LCT, examined in detail several distinct disciplinary forms of knowledge and knowledge-building activities, consequently characterizing these as *internal conceptual languages*, or frameworks of knowledge. These language are 'internal' in the sense that they specify precisely those forms of knowledge constitutive of a discipline, i.e., the categories, structures, and conceptual relationships with which theories within the discipline are expressed and developed. Such organizations include both the specific terms and relations between terms practiced in a discipline and any more generalized structuring systems (metaphors, paradigms, and so on) that provide a sense-making background for those terms and relations. Bernstein argued subsequently that simply having a sophisticated internal theoretical language for disciplinary reflection of this kind is not enough for the effective development of knowledge. The internal language must also engage with something 'outside' the theoretical descriptions produced. In particular, the theoretical language must engage with the *body of phenomena* about which a theory is intended to be a theory.

Bernstein considered this second component of knowledge building as a task in its own right that had not so far received adequate attention. Most disciplines he examined had failed to thematize the mechanisms necessary and, as a consequence, were "deaf to data" (Maton & Chen 2016: 29). Bernstein (2000: 29) characterized this mismatch between powerful internal conceptual languages and relatively weak descriptions of relevant data as a *discursive gap* and argued that explicit engagement with the discursive gap is an essential precondition for effective knowledge-building. Whenever this is not done, disciplines have a tendency to impose "theory onto data in a 'cookie-cutter' model which ignores the particularities of objects of study" (Maton & Chen 2016: 29). It is telling that this is also a rather exact characterization of several critiques that have now been brought against certain branches of multimodality research (cf. e.g., Forceville 1999; Ledin & Machin 2019).

To overcome the discursive gap, Bernstein articulated a further meta-theoretical distinction between a discipline's internal conceptual language and *external* 'languages of description' that serve precisely the purpose of organizing objects of investigation, i.e., data, in ways that make those objects accessible to analysis. Crucially, these external languages offer ways of characterizing data *without already enforcing internal theoretical distinctions on that data*, since this may well turn out on subsequent investigation to have been inappropriate or premature. External languages of description are intended to address this problem directly by allowing characterizations of data that can always go beyond what might be predicted by a theory. Only in this way can the finding of counter-examples or new phenomena be supported, which can in turn lead to a theory being forced to accommodate new empirical results. In general, therefore, it falls to a discipline to define such external languages of description to explicate the methodological and practical steps that negotiate appropriate relations between theory and data. This involves establishing relations that do not simply impose theoretical categories on data, but which instead are open to the specificities of data without losing sight of the more general conceptual goals and frameworks that a discipline or theory is pursuing. In many respects, this can be seen simply as an attempt to be rather more detailed about what it is to 'apply' a theory at all.

Here it is particularly important to emphasize that this definition of external languages of description in no way assumes that data can be approached as if the analysis is theory-free. Data will always be being viewed from a particular disciplinary perspective and below, when we introduce the characterization of materiality that we will employ, it will be seen that theoretical questions and the particular conceptual organization of internal languages of description remain the primary motivating orientation. The principal contribution to methodology then lies in the relative reliability in application that any external classifications provide for organizing data and operationalizing theoretical distinctions. In certain respects, preparing access to data in this way may be likened to the preparation of corpus data in linguistics by employing *relatively* neutral, or low-level, annotations upon which more theoretically developed characterizations can be built and hypotheses tested. Just as even the most 'neutral' linguistic annotations are embedded within a broad understanding of potentially useful linguistic distinctions, this is the case for our account of materiality as well: the account remains strongly semiotically motivated throughout.

3 Construing Materiality as an External Language of Description for Multimodality

In the previous section it was explained how, according to the constructs developed within LCT, an external language of description makes explicit how phenomena can be organized most effectively for addressing particular theoretical concerns, but without already assuming that the analytic categories of some theory or discipline are fixed or unproblematic in their application. In the evocative phrasing of Gunther Kress, the question addressed by this task is precisely that of how to turn 'stuff' — the material and phenomena with which one is directly confronted — into 'data' — a sensibly organized array supportive of systematic analysis (cf. e.g., Kress 2012: 252–253). As indicated above, we have suggested elsewhere that the characterization of materiality articulated by Bateman et al. (2017: 101–110) might fulfill precisely this function and so serve as an appropriate external language of description for multimodality research. This proposal will now be explored in detail.

Particularly in areas of more explorative multimodality research, the sheer variability and richness of the phenomena at issue can readily overwhelm researchers, particularly those who are newer to the field. Whereas more established disciplines or disciplinary areas, such as linguistics, already have strong methodological guidelines concerning how to approach data, this is not the case for multimodality research. Researchers embarking on empirical analysis are then often left to their own devices concerning just what might be relevant and what not. This challenge may even speak against attempting empirical analysis at all — if it is unclear just what phenomena are relevant, then empirical analysis can readily appear premature. Bateman et al. (2017: 213–221) argue that explicitly engaging with the kind of materiality involved in any multimodal communicative situation under investigation can help reduce the overall complexity of this problem by forcing 'natural' divisions on the objects of analysis. 'Natural' is meant here in the sense that the materiality involved will itself reveal organizational structures that can beneficially guide analytic attention and support principled decisions concerning just what must be included within analysis and what not.

3.1 The Initial Classification

The empirical methodology defined by Bateman et al. (2017: 230) sees the first stage of analysis as one of organizing the object of investigation according to the materiality involved, classifying that materiality along several dimensions. Four of the basic dimensions of materiality are shown graphically in Figure 1: temporality, space, role, and transience. These dimensions should not in themselves be particularly contentious as they echo several existing proposals for classifications of material possibilities made in the literature, including both work in studies of multimodality and theoretically distinct explorations of the nature of 'medium'. What is new here, however, is the explicit construal of these dimensions as an external language of description for empirical multimodality research.

All of the dimensions and their further refinements discussed below are concerned essentially with properties of the *traces* that materials are capable of supporting when used for communication. This is what makes the account a semiotic account rather than one that is simply classifying 'material' as such. What the dimensions are intended to cover can be specified briefly in the following terms, running left to right in the figure. First, the dimension of temporality captures whether the traces supported by a material can themselves change over time (dynamic) or not (static). For example, a photograph can deploy only static traces, whereas the material created by film supports traces that move of themselves. Second, the dimension of space captures whether traces are either two-dimensional, i.e., flat, or three-dimensional, i.e., extended in depth. This is independent of what the material is being used for in any specific case: thus, for example, a flat monitor display showing a three-dimensional building is still 2D, and so on. Third, a material can either support an 'observational' role for any interpreter — i.e., the interpreter is inherently 'outside' of what is being interpreted, as is the case with a painting — or a 'participatory' role for any interpreter — i.e, the interpreter is *inside* and, indeed, 'part of' the material being used. The latter is the case, for example, in face-to-face natural conversation or in a virtual reality computer game. Further discussion and examples of this dimension are given in Bateman et al. (2017: 96–99). Lastly, the traced representations or depictions can vary with respect to whether they are simply present and stable ('permanent'), or whether they appear and disappear, each potentially being replaced by the next ('fleeting'). Speech, for example, is fleeting in that each sound is immediately replaced by the

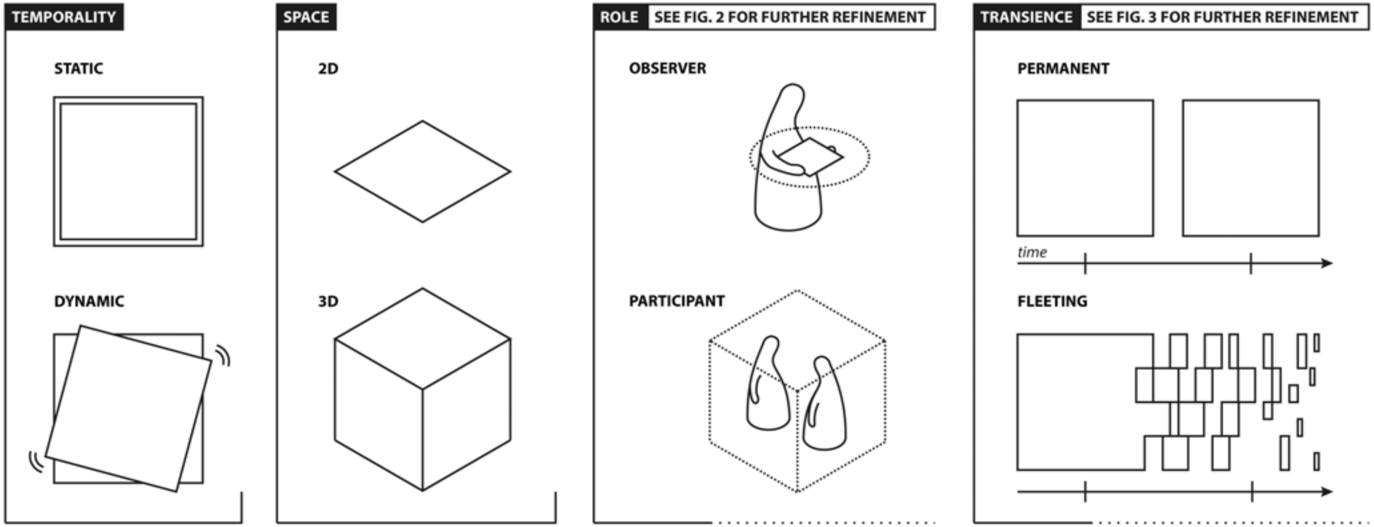


Fig. 1: Four basic dimensions of materiality adapted from Bateman et al. (2017: 109). Visualization: Jana Pflaeging.

next, whereas printed text is permanent. There are a range of more fine-grained distinctions that can be made here which will be seen in more detail below.

When addressing any multimodal communicative situation, the proposed methodology begins by attempting to find the most appropriate placement for that situation according to the four dimensions defined. As a simple example, the pages of a traditional comic book are static: they do not support representations or traces that themselves move, in contrast to film or video, where representations or depictions that themselves move are readily supported. Since this is a question of the possibilities of the materiality at issue and not what is being done with that materiality, it is of no consequence if a video happens not to include movement: this is still a *possibility* of the material. Note here again that it is the material as presented to and for perception that is crucial — that is, it is not the (in earlier times) celluloid strip of semi-transparent material that constitutes the semiotically-relevant physical reality of film but rather the (virtual) materiality of moving forms of light and shade (and sound) constructed by such strips and their corresponding machineries of projection. This then already includes features related to cutting, editing, framing, etc. — i.e., technologically supported possibilities. Taking these properties together results in a detailed characterization of the affordances of the medium. The *relevant* materiality is consequently what is made available for perception as this is the sole source of semiotically significant distinctions.

Moreover, as Bateman et al. (2017: 218–220) illustrate, communicative situations can, and often do, exhibit highly complex material structures with distinct levels of embedding, each of which then *requires its own material classification*. If this complexity is not made explicit prior to commencing analysis, the most likely result is confusion — both on the part of the analyst and the resulting 'analyses'. The act of classification itself can be used to guide this process: whenever the analyst would otherwise be forced to apply mutually inconsistent classifications, this is evidence that different levels, or 'slices', of analysis need to be assumed. Indeed, whenever classification appears difficult, this is generally a strong indication that additional structure is necessary. As an example, a material cannot be both two-dimensional and three-dimensional and, as a consequence, should these distinctions appear to co-occur within some object of investigation, then it can usually be postulated that the communicative situation must in fact be divided into at least two slices: one exhibiting a three-dimensional materiality and a further, often embedded, one exhibiting only two-dimensional materiality. A straightforward case where the classification enforces slicing would be the multimodal situation of a caregiver reading out of a picturebook for a young child: the situation overall is clearly three-dimensional, but within that there are the various two-dimensional slices of the pages of the picturebook. The latter are embedded within the former. Another example, rather more complex, is the situation where one might be analyzing a narrative film. Here one might ‘zoom out’ so that the communicative situation at issue is the actual reception situation of a cinema with an audience, seated in particular configurations and distances from a screen, with particular lighting and sound conditions, and so on; or, we might conversely ‘zoom in’ so that it is ‘just’ the two-dimensional on-screen audiovisual depiction that is the object of study.

Analysis can only proceed when it is made quite explicit with respect to which slices questions are being addressed. Each analytic slice may bring different materialities, temporalities, and semiotic modes into consideration. On the one hand, this encourages ‘activity’-based approaches to analysis as promoted for multimodal studies of interaction by Murphey (2005), Norris (2009), Bucher (2011), and others. On the other hand, it also naturally includes a notion of ‘variable analytic focus’ which allows the analyst to move freely between object and performance as targets of analysis. For larger-scale empirical work, there is also a direct link to be drawn between appropriate slicing and appropriate levels of corpus annotation: since each slice may have different materialities (and hence different temporalities, affordances, and accompanying semiotic modes), it may require distinct analytic procedures and corresponding coding schemes.

3.2 Material Refinements

There are several interdependencies between the dimensions that need to be considered before proceeding: that is, it is not the case that situations can be positioned with absolute freedom along each of the dimensions given. This is particularly the case for any materialities semiotically engaging ongoing experience since this demands a temporal unfolding within which the experiencing can take place. Moreover, once there is experience, then this needs to be mediated by some perceptual system(s). As a consequence, encountering materiality via the ‘participant’ pole of the role dimension is going to engage participants and materials in certain specific ways, although the details of how distinct materialities do this may vary for different semiotic modes. For example, for a musical performer using some musical instrument there will be (at least) acoustic, haptic, and motoric engagement. For someone watching and hearing the performance, the haptic and motoric components will not be present, although if they are performers themselves there may well be resonant neural responses. This latter phenomenon plays a crucial role in many media but has probably been most extensively discussed in relation to film (cf. Sobchack 1992; Barker 2009; Gallese & Guerra 2020).

In addition, the senses will be engaged, in a variety of combinations, via the ‘observer’-pole of the role dimension as well. In such cases, one is not participating in a situation involving the senses but examining/observing a materiality from a distance, separate from that materiality. As long as the materiality is being manipulated for semiotic purposes so that the distinctions in perception are being actively shaped, then this is sufficient to motivate including it in a semiotically-inflected view of materiality. A good example of use of an observer-oriented haptic materiality would be a tactile map for those with sight difficulties: the user of such artifacts is not a 'participant' in the information communicated by the map. The same can be said theoretically for the other senses as well, although these are clearly very different in the ranges of deployment found: observer-oriented smell might be considered in the context of perfumes, while observer-oriented taste might potentially be invoked by a semiotics of food. Even though the 'perceiver' is evidently participating in perception, this is not the locus of semiotic traces relevant for constituting the communication and so is clearly observer-oriented; there are also parallels to be drawn here with consumer/producer and performer/audience relations. Figure 2 summarizes this refinement of the role dimension graphically.

The actual physical act of *producing* traces in materiality can itself also provide possibilities for semiotically-significant variation which consequently need to be addressed in the model. For example, not only can materials 'push back' when traces are inscribed in them, but different materials will push back in different ways. In the general case, each particular semiotic mode requires its own particular kinds of traces when realizing, or materializing, its technical features and, consequently, not only do different materials push back differently, but they may also push back differently depending on the *particular kinds of traces that any given semiotic mode needs to inscribe*. This relies on the crucial distinction between a materiality 'as such' and the particular subsets of material dimensions constituting the *canvas*, to which we return below. As a purely physical analogy, for some particular use of wood it may be sensible to 'go with the grain', so that breaks are avoided; but for that very same material, for some other use it might be more appropriate precisely to go against the grain, which would make very different properties of the same materiality apparent. This is little more than the general statement that how a material responds will depend on what you do with that material, but it is a crucial distinction in order to pull apart materiality 'as such' and what happens when that material is used semiotically.

This particular aspect of multimodality can almost be considered the 'obverse' of affordances and is now receiving attention from various perspectives, ranging from attempts to characterize 'touch' within sociosemiotic accounts (e.g., Bezemer & Kress 2014; Jewitt 2018) to the use of 'force feedback' in an increasingly diverse range of human-computer interaction devices; Jewitt et al. (2020) overview recent developments in the field, addressing potential social implications in particular. The positive value of such feedback for meaning-making has been the subject of

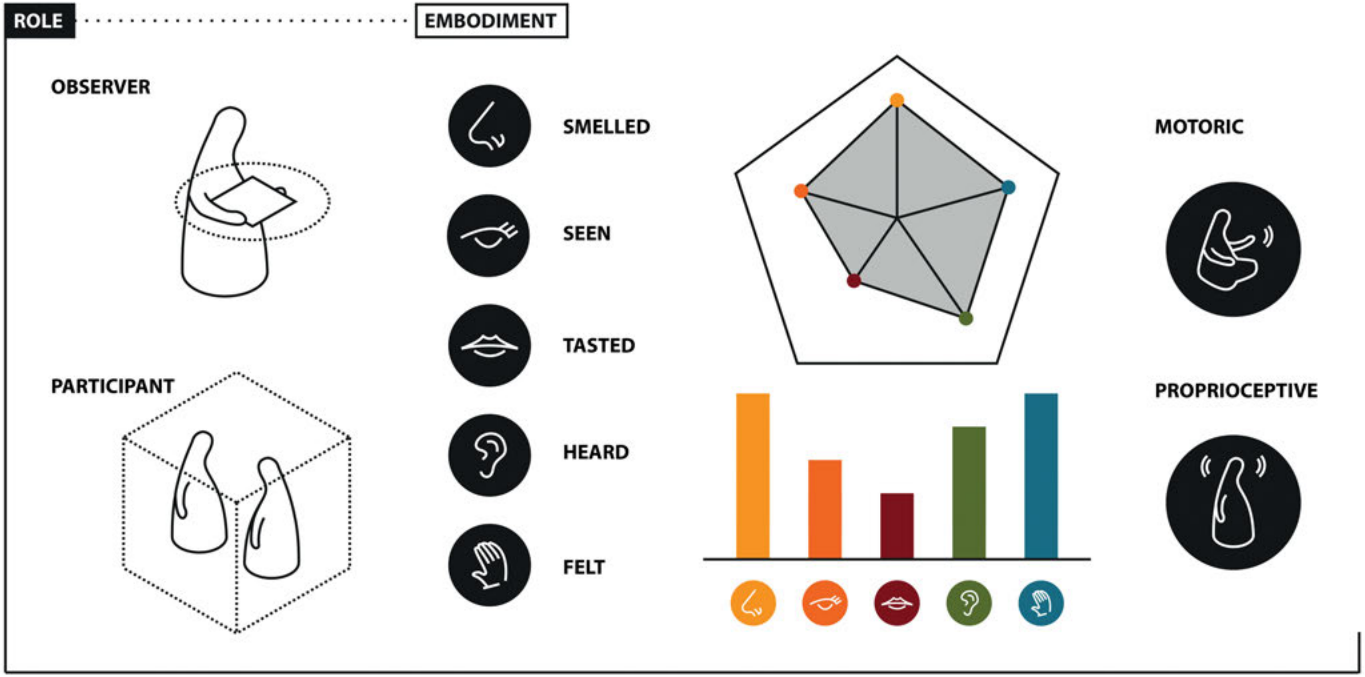


Fig. 2: Refinement of the basic dimension of role to incorporate sensorial engagement between material traces and corresponding participants and observers. Visualization: Jana Pflaeging.

many empirical studies in human-computer interaction and, from the present perspective, it is now clear that accounts of materiality must be sensitive to this dimension of variation as well. The view of materiality being developed here makes this relatively straightforward because materiality is considered a site for 'trace-leaving' in a completely general fashion. The radical multimodality of traces is accepted as fundamental for any full account of multimodality. This does not then distinguish between the kinds of 'traces' carried by haptic material regularities and more traditional notions of traces as marks or sounds.

Resistance, or push-back, from materials is closely entwined with another of our basic dimensions: the dimension of transience. We can, therefore, usefully 'unfold' this dimension further as well. (Im)permanence is not, after all, a simple continuum. First, the manner of appearance of traces may vary: they might appear instantly or take time to appear; conversely, the manner of *disappearance* of traces may similarly vary. For this reason, we can additionally characterize the permanence-fleeting dimension in terms of variations in *the rate and manner* of appearance-disappearance or, to adopt terms from music and sound design, 'attack-decay'. If materialities support such variation, then the shapes of those variations may well be taken up for semiotic purposes — that is, this variation will be available for corresponding semiotic modes. Second, variation may also occur at different granularities. For example, speech is highly transient and so is only fleetingly available, whereas a slide show exhibits a far lower granularity: each slide is shown for some period of time (during which the material may appear permanent) and only then is replaced by the next slide. For the former case of speech, the granularity involved is the continuous stream of sounds; for the latter case, the granularity might be seen on a 'per slide'-basis and the various options for proceeding from slide to slide (e.g., sudden transitions *vs*. fades) correspond to variations in attack-decay. And third, we must also consider an entirely different kind of permanence or persistence that is supported by embodied perception: if, for example, someone makes an upwards spiraling gesture in the air, the visual trace disappears as it is being made (transient, high granularity), *but* there will generally still be a clear 'motor memory' of the gesture and its dynamic unfolding (cf. Mittelberg 2017; Schüller et al. 2017; Tversky & Kessell 2021, this volume). A similar phenomenon is evident in speakers' of Japanese disambiguation of homophones by 'writing' the corresponding character on the palm with a forefinger.

Such 'traces-in-memory', in all likelihood supported by an inherent linking of perception and motor mechanisms, demand a more nuanced account than considering them simply as transient visual experiences would provide. This additional route by which relative transience can be manipulated we term the 'time-depth' of the materiality. A greater time-depth often corresponds to the persistence of *spatial* information concerning any traces formed — a property of (our engagement with) materiality that clearly plays an important role in sign languages, where both movement and positions-in-space are highly semiotically charged and, often, persistent beyond their bare performance. These three further refinements of the transience dimension are summarized graphically in Figure 3.

The extensions to materiality captured by these refinements emphasize again how materialities are always entangled with (embodied) perception. This significantly changes the kinds of manipulations that are possible and suggests many further intriguing theoretical questions. For example, might the material 'persistence' of audial 'space' extend not only into the past but also *into the future*, such as when continuations of note sequences or melodies are predicted? This would make triangulations with phenomenological studies particularly interesting. Alternatively, moving in quite a different direction, Pirini uses varying degrees of material transience in order to motivate a highly multimodal account of 'intersubjective materiality' (Pirini 2016). In this framework, actions unfolding within nested materialities that exhibit 'fleeting', 'adjustable', or 'durable' degrees of transcience offer a potential bridge from micro-scaled activities to more extended (and durable) understandings of intersubjectively shared situations. There is evidently very much more to consider here.

3.3 The Core Distinction Between Material and Canvas

Finally, for this initial characterization of the possibilities attributed to materiality, we need to explicitly relate materiality to its more general anchoring within the multimodality framework assumed. As noted at the outset, the characterization of materiality pursued here is not that of physics but rather rests on active perceivers' embodied engagement with materials for semiotic purposes. We need in addition, therefore, to say more about how the view of materiality and semiotic purposes relate. That is: it needs always to be borne in mind that the distinctions set out are anchored in the possible uses of material for semiotic purposes, i.e., for communication in the very general sense assumed by multimodality.

The use of any materiality for semiotic purposes necessarily involves 'inscribing' or 'shaping' material-perceptual entanglements in particular ways. In the model set out in Bateman et al. (2017: 86–87, 114), the entailed mutual relationship between material, on the one hand, and the particular traces that a given semiotic mode requires, on the other, is captured by the theoretical construct of the *canvas*. The canvas of a semiotic mode is simply that semiotic mode's materiality *when viewed with respect to the specific forms of traces required by that semiotic mode*. The notion of canvas consequently captures the inherent linking of form and material that is definitional for each and every semiotic mode and also makes it clearer how

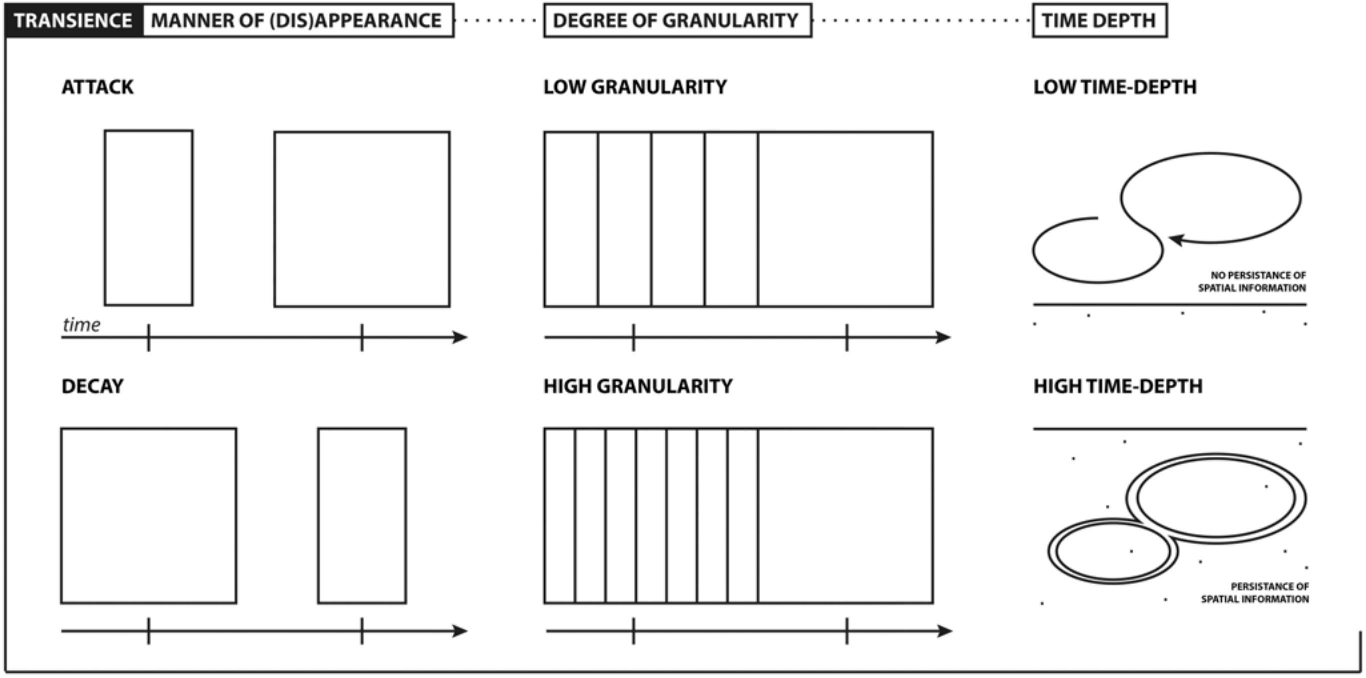


Fig. 3: Refinement of the ‘fleeting’ pole of the basic dimension of transience to incorporate the three distinct kinds of variation in the temporality of such traces. Visualization: Jana Pflaeging.

‘single’ materialities may readily support multiple semiotic modes at the same time. From a production or design perspective, the canvas is what the sign-maker has to work with (or against: see above) when traces for a particular semiotic mode are to be produced. Thus, alternatively expressed, the canvas designates the role that material plays when mobilized at the lowest level of semiotic abstraction in the overall tri-stratal model of semiotic modes developed in Bateman et al. (2017: 113–117).

The consideration of materiality ‘through the lens’ of its deployment as part of a semiotic mode is then central to the definition of ‘canvas’. This distinction is beneficial for several further related theoretical and descriptive challenges that commonly come to hinder empirical research. It is also important when considering semiotic relationships across different materialities and media. Precisely because the canvas of a semiotic mode characterizes just those material distinctions or regularities that are necessary for that specific semiotic mode to operate, one can productively view the canvas as an ‘abstract’ or ‘generalized’ materiality as well (Bateman et al. 2017: 103). This means that, to the extent that any materiality supports at least the material requirements given by the canvas of some specific semiotic mode, then that materiality may serve to materialize the technical features of that semiotic mode. The material is then capable of supporting the traces that the semiotic mode demands. This extends naturally to a reconstrual of transmediality, i.e., the use of semiotic modes across different media. At one and the same time, the model insists on the inseparability of semiotic modes and materiality and yet still allows semiotic modes to occur with different materialities, such as paper and screen. What must remain constant across the different materialities are *the particular properties identified by the canvas*. As long as different materialities can play the same canvas role vis-à-vis the semiotic mode, the minimal material constraints for the use of the semiotic mode are met, because the particular perceptual experiences necessary for recognizing and using the semiotic mode are supported. Thus, in short: canvas properties explain why it is equally possible to ‘read text’ on paper, a screen, spray-painted on a wall, or carved in stone; however, canvas properties do *not* (on their own) explain variants such as written and spoken language — for this, more complex interrelationships are required both between materialities and across other components of the semiotic modes employed.

To summarise, then, the dimensions that have now been given are not intended as characterizations of material as such, although clearly there are some dependencies. It is not the material itself that is relevant, “but rather the *semiotic construction of access* to that material" (Bateman et al. 2017: 101). The central concern of this characterization of materiality has therefore been to provide a formal account of how material properties, or affordances, have consequences for the ability of materials to carry ‘traces’. It is the traces (and, inseparably, their perception) that are key: practices of communication can only become conventionalized and develop as semiotic modes via such perceptible traces, which themselves may *simultaneously* span and engage any and all sensory channels. This means that, although we are concerned with materiality only as it becomes accessible for semiotic purposes, it is equally inappropriate to see this primarily in terms of sensory channels. In fact, we see the focus of materiality as presented here as an important way of superseding earlier approaches working more directly in terms of sensory channels, such as, for example, distinguishing between visually carried information and aurally carried information. Such approaches generally fall foul very quickly of inherent entanglements between materiality and that materiality's semiotic use and, as a consequence, come in any case to make additional distinctions on evidently semiotic grounds. Even accounts that explicitly orient towards materiality and its centrality have tended previously to fall into this conflation of the material and the semiotic, making analysis and demarcation of data unnecessarily complex.

4 Examples of Different, but Related, Materialities in Communication

The dimensions of materiality set out in the previous section are considered essential for relating any phenomena of interest to their systematic analysis. Moreover, as an external language of description for multimodality, they need: (a) to be supportive of analysis without assuming the theoretical constructs of that analysis already hold; and (b) to be directly applicable to phenomena of interest in a manner supportive of reliable application and operationalization. However, although their characterization in the previous section is straightforward in principle, their precise implications and manner of application to support concrete empirical analysis requires practice. In this section, therefore, this application of the account of materiality as a link to data will be illustrated by means of a collection of different, but nevertheless semiotically related, communicative situations. Examining the situations from the perspective of materiality will be shown to suggest webs of connections between traditionally rather different communicative situations as well as making the situations themselves more amenable to analysis.

4.1 From Blackboards …

Let us assume as a starting point that we wish to analyze the use of blackboard and chalk as a teaching aid in a classroom or lecture situation as is, in all likelihood, all too familiar. The first question to be asked is then how the materiality of this situation is best to be placed within our overall framework. This has been argued above to constitute an appropriate methodological step for preconfiguring the kinds of semiotic modes that might be relevant in subsequent analysis.

To begin, a blackboard including its content is evidently two-dimensional as the information it presents (assumed to be carried in the various forms inscribed on the blackboard's surface) is distributed around the blackboard in more or less recognizable, i.e., segmentable, spatial regions. We might note that visually there appears to be a possibility of varying color somewhat as marks may be distinguishable in this respect and that usage may appear (only analysis will confirm or deny this) not to be random. Marks made on the blackboard will be static and we are clearly detached from it, i.e., we are external observers. So far there is no particular difference between this material and, for example, the pages of this book: we might then encounter similar semiotic modes and so empirical results concerning those semiotic modes in one communicative situation might be expected (or, better for empirical subsequent study, predicted) to apply to the other.

Considering this further, however, there may well be aspects of the description offered so far that can strike us as insufficient. Considering the dimension of material permanence/transience, we might ask, for example, whether the marks are really permanent, as they are in a printed book. Is it not the case that the chalk marks might be erased? Is that not one of the motivations for using a blackboard in the first place? Paying attention to this line of reasoning is not compatible with the characterization given above and so we are *forced* to reconsider the exact slice of the communicative situation and its materiality that is relevant. This is precisely the function that we want for our account of materiality as an external language of description: it should guide us to particular characterizations of the communicative situation rather than others so that the basis for further analysis is made more secure. In a genuine blackboard scenario, it is not the case that things on the blackboard disappear by themselves; nor is it the case that the things on the blackboard appear by themselves fully formed either — they are written on the blackboard and, perhaps, subsequently erased, adapted or replaced *by someone writing on the blackboard*. This situation is neither static nor two-dimensional and so we have a first indication of the necessity of structuring our communicative situation into further slices.

This means that we have a three-dimensional, dynamic situation in which someone is making marks on a blackboard. This situation is transient but still observed/external/detached. Situations of this kind support a range of quite different semiotic modes, including, various movements of the writer — such as pointing, waving, gaze, and other gestures that might not have been evident in the previously discussed slice at all — as well as, if we do not restrict ourselves to the visual sensory channel, spoken language. In considering an extension of the materiality at issue in this way, the question should always be whether *material traces* are being formed or shaped that show sufficient regularities to make semiotic interpretation beneficial. This is not, therefore, simply extending the material circumstances arbitrarily as occurs to us, but rather always an extension precisely because there appears to be a shaping of material traces in the extended materiality that renders semiotic interpretation likely to be productive.

Thus, for example, we might consider an analysis that takes the perspective of the person writing on the blackboard, thereby shifting from an observer materiality to a participant materiality. This is analogous to any shift we make when analyzing a semiotic activity from the performers' perspectives rather than an observer's. Whether this is a useful thing to do itself depends on just how much meaningful regularity and variation is being brought to that performance. Whereas, this may have enormous weight when analyzing the semiotic practices of a dancer or the actors in a theatre performance, it may not prove particularly revealing for the blackboard situation — *unless* the person writing on the blackboard is bringing particular performative flourishes to that activity! Discussion here is relevant for considerations of the source of attributions of artistic contributions across media such as theatre and film, such as Carroll's (1996: 69) classification of theatre as a performing art but film (or the moving image) not; Tseng (2017: 130) also discusses this further.

More commonly, the situation of writing on the blackboard (possibly accompanied by spoken language) is not used as an opportunity for artistic creativity and so can be limited to two essential material shapings: one is the temporal shaping, segmentation or 'layouting' that divides the stream of activity into temporal parts, the other is the 'semi-permanent' visual trace on the blackboard. These are, because of their common origin within the considered slice of materiality, intimately (and formally) related. That is, although the two-dimensional traces on the blackboard are themselves clearly static traces — they cannot move — they have a very specific history of dynamic production and are, as a consequence, necessarily entangled with that history. In short, the static traces on the blackboard may also be read as an external record (index) of how they unfolded in time.

This is, in fact, often essential: most of us will be familiar with the state of a blackboard after a lesson: the information on the blackboard might well appear nothing short of chaotic, but when placed in the temporal context of its production can be seen to have been developed in a logical and self-explanatory manner. It is only when the traces on the blackboard are considered severed from their history of production within the embedding communicative situation (and its materiality) that that order collapses — or, rather, is no longer accessible.

Writing on the blackboard thus shares semiotic features with a host of other situations where the temporal unfolding of the production of external traces plays a crucial role, ranging from giving someone verbal directions while drawing them a map to architectural design, which may actively imagine space drawing freely on visual traces, objects, language, and gesture (e.g., Murphey 2005). Again, these material commonalities will suggest areas where similar semiotic modes, and the results of those semiotic modes' empirical study, may be applicable. Methodologically this can then play an important role in motivating and predicting extensions and transfers of empirical results across quite diverse situations.

One can also explore variations in communicative situations more systematically to show where there may indeed be different semiotic modes involved. If we were, for example, to consider instead a whiteboard and whiteboard markers, this is very similar to the blackboard but still, nevertheless, exhibits a few differences in materiality that could under certain circumstances have semiotic import. Marks on the surface of the whiteboard can (in the ideal case) be erased with considerably less effort than is the case for the blackboard. This may have consequences for the kinds of uses made of the two materialities — use of the blackboard may avoid the need to erase more than is the case for the whiteboard, resulting in a treatment of the whiteboard situation as slightly more 'transient' than the chalk-on-blackboard situation. The granularity of persistence would be broadly similar in both cases, however.

4.2 ...to PowerPoint Presentations ...

Going further, if we were to change our target of analysis from writing on a blackboard or whiteboard to producing presentation slides with a presentation program such as PowerPoint or Keynote, much more variation is evident: there are clearly similarities in materiality (which would predict similarities in the semiotic modes available) as well as significant differences. In this case, probably the clearest difference is that there is no longer any entanglement with the history of production: this is no longer accessible and cannot be semiotically shaped to enrich the two-dimensional material traces of the slides in any case. This places greater weight on other techniques and material potentials to support interpretation of what is shown — most particularly on the deployment of two-dimensional spatial layout to guide reception, thus aligning again more with printed materials, generally accompanied in actual situations of use by spoken language (and, if one can see the presenter, gestures as well). The latter situations overlap with the blackboard situation and so again share semiotic possibilities; the former may of course be employed on the blackboard as well, depending on how the writer chooses to structure the visual traces, but is not necessarily present.

The materiality of the slide presentation also allows, however, a *depiction* of the temporal unfolding of the blackboard case — in particular by successively introducing traces onto the projected slide, changing those traces, and subsequently removing them. This is not the same as the blackboard case, as the appearance or disappearance of traces is not linked to an actual production situation: it is instead part of the temporal constitution of the materiality of the slides themselves. These are transient, but with a far lower 'transience granularity' than is the case, for example, with speech. Indeed, one of the capabilities specifically of computationally supported media is a kind of *generalized transience* (Bateman et al. 2017: 100), where the 'permanence' or not of traces can be arbitrarily varied. There is much to research here, but for present purposes a notion of semi-permanent, low granularity transience will suffice, standing in the present case as a (metaphorical) depiction of a production situation which serves the role of entangling (generally) static visual traces with a history of production. This directs attention, guides interpretation and may, for more complex presentations, serve precisely the additional organizational role discussed for the blackboard case.

4.3 ...to Sand

Our last related communicative situation turns to a very different body of multimodal practice, the audiovisual story-telling tradition of 'sand stories' prevalent among speakers of the Arandic languages in Central Australia. In particular, the focus here will be on Green's (2014) detailed empirical analysis of this practice from an explicitly multimodal perspective. The communicative situation of telling sand stories generally consists of a story-teller seated on a sandy patch of ground who narrates using spoken language, gesture, elements of sign language, small movable objects such as twigs and leaves, and combinations of conventionalized marks and iconic (diagrammatic) traces left in the sand. Much of the communication is then carried visuokinetically, where it is not only the product of the traces left in the sand that is relevant but also their manner of being made as these unfold in time.

To analyze these complex ensembles empirically, Green proceeds by making multileveled corpus annotations of a collection of video recordings of sand stories being 'told' in their natural environment. Green's descriptions of both the theoretical decisions made and the practical challenges of annotating such a richly structured semiotic ensemble are extremely valuable in their own right, demonstrating well the issues that occur. Moreover, despite the considerable apparent differences between this communicative situation and those, no doubt more familiar, situations we have addressed in this chapter so far, significant overlaps in the materialities involved reveal that most of the concerns that Green encountered should also be seen to be of concern for these other situations as well. And, in addition, of particular relevance is the reoccurring challenge faced by Green that the signifying practices examined cut against many of the lines of demarcation traditionally assumed in multimodality research.

Thus, gestures that begin as hand and arm movements in the air frequently turn to movements that leave traces in the sand, continuing their formerly purely 'virtual' trajectories; similarly, movement traces in the sand can leave the ground and continue their paths in the air. Story elements may be introduced verbally, only to continue in the movement of objects across the sand. And there may even be gestured spatial references to traces and objects *that are no longer physically present*, having been erased previously by wiping the sand clean. This flexibility led Green at the outset not to use the term 'semiotic mode' but instead to adopt a broad notion of 'semiotic resource', which, in the multimodal tradition of Kress & van Leeuwen (2006 [1996]: 9–11) and others is taken to emphasise semiosis as a means of addressing communicative tasks rather than as being a rule-bound deployment of more or less fixed sign inventories (cf. e.g., van Leeuwen 2005: 3–6, 285; O'Halloran 2009: 98; Kress 2010: 6–7).

Many uses of 'semiotic mode' in the multimodality literature are indeed unhelpful for empirical analysis just as Green describes because they rest primarily on open-ended lists of examples — e.g., writing, gesture, body, color, laughter, 'and so on' — in lieu of definitions. This practice, criticized in some detail in Bateman (2019a), is still very common in multimodality research and results in 'semiotic mode' saying little more than is already covered by the term 'semiotic resource'. Green consequently reserves the term 'multimodal' for multisensorial communication, similarly to the approach of Fricke (2013) and many others.

As Green's analysis proceeds, however, it is unclear whether the shift to talking of 'semiotic resources' so as to avoid the pitfalls of potentially inappropriate boundaries between poorly defined 'semiotic modes' is effective. As Green at one point in the discussion explains:

> Some moves leave a mark on the ground, others are enacted in the air and others are deployed in both media. My decision to code *moves* according to the medium in which they are enacted — on the earth, in the air or in a combination of both — forces us to look at visual data in a new way. It does not presume that there is a clear-cut distinction between various semiotic systems and provides a framework for comparing drawing and gesture which highlights both their similarities and differences. [...] there are times when a drawing movement becomes airborne and hence from one perspective may be regarded as manual gesture. There are also

> times when gestural movements in the air contact the ground and leave a mark. (Green 2014: 79)

Considerable theoretical uncertainty therefore remains concerning just how potentially ‘overlapping’ semiotic systems might best be approached, both theoretically *and* practically during analysis. This is not helped by the fact that the notion of ‘semiotic resource’ is also intrinsically vague — anything that may serve a semiotic purpose may be a resource: van Leeuwen even writes, for example, of ‘genre’ being a semiotic resource (van Leeuwen 2005: 128). This does not provide support for empirical analysis. Moreover, even Green’s use of ‘media’ in this quotation is clearly reaching towards notions of materiality, but little guidance for driving the empirical analysis further is secured precisely because there is no formal characterization of just where similarities and differences in such materialities might lie.

4.4 Building Connections

Crucial for the present discussion is then the fact that the issues that Green raises with respect to the analysis of the sand stories are far from individual cases. Indeed, as Green herself notes (Green 2014: 238–240), phenomena of these kinds are widespread. We might well imagine someone ‘drawing’ a map to some destination by tracing the path in the air; while a (albeit rather limited) set of in-air gestures might well accompany chalk marks being made on a blackboard, probably adding connotations of continuation, lack of boundaries, approximation, or emphasis. Such constructions extend relatively straightforwardly to any communicative situation where there is a rich and manipulable environment available for semiotic shaping and thereby make contact with formerly rather distinct areas of discussion, such as the use of gesture in scientific discourse (e.g., Roth 2000) or closer considerations of the relationship between gesture and diagrams more generally (e.g., Engle 1998; Kang et al. 2015; Tversky & Kessell 2021, this volume).

A stronger basis for comparison and transfer of insights, results, and questions across domains is offered by an appropriate characterization of materiality. It is not the ‘medium’ distinction between a trace in sand (or on a blackboard) and a path in the air that is criterial but rather the common canvas that these two facets of the material situation share, or rather co-construct. Both sand (or blackboard) and air, i.e., some visually accessible region of space, support dynamic, observer-oriented traces and it is these traces that offer a single canvas for corresponding semiotic modes capable of spanning the distinct materialities involved. Figure 4 summarizes the discussion by showing three of the example situations discussed in this section side-by-side. This illustrates both the commonalities in materialities and overlaps in potential semiotic modes that those materialities support as the communicative situations become progressively more complex. The labels for potential semiotic modes in the figure are, as always, to be read as abbreviations for ‘meaning-making mechanisms’ making use of the identified resources: that is, meaning-making with spoken language, meaning-making using sketches, menus or mouse-clicks, and so on — whether such semiotic modes are actually operative is, as always, an empirical question. Thinking about communicative situations in this way makes it very clear just how explicit one needs to be about *what* one is focusing on. For example, in the sand story situation the slices depicted in the figure concern, first, the participating performer-and-audience slice, including body posture, gesture, proximity, speech, and so on, and second, a strictly dependent slice that includes observing just the traces being left in the sand. Similar distinctions are drawn out for the other two situations shown as well.

There is considerably more to discuss here, particularly with reference to the distinct kinds of relations that can hold between the distinct material ‘slices’ involved in each situation. In the figure each of the three communicative situations is shown as involving two material slices with a corresponding relation holding between them. The relations give rise to progressively more semiotic ‘distance’ between slices as we move from left to right. In the sand story situation, the relation is one of ‘enhancing’, where a material slice remains an integral component of a larger slice but enhances that larger slice with the possibilities of a limited spatial record. In the blackboard presentation situation, the overall configuration is similar, but the blackboard acts as a far more fine-grained addition capable of recording utterances, spoken-about diagrams, and so on: here, consequently, the relationship is seen as one of ‘projecting’. Lastly, in the situation of *creating* a PowerPoint presentation, the relation is one of ‘designing’, where the activities of one slice give rise not to another slice directly, but rather create an artifact that may function as, or give rise to, a material slice for incorporation in a quite distinct communicative situation (e.g., most commonly, a PowerPoint presentation). Thus, the sand stories are relatively distinct from the other two cases since the 2D-space of the sand is maintained as part of the overall performance space — reminiscent of what Rosenberg (2008) refers to as ‘ideational drawing’ or a ‘thinking space’. Aspects of this potential may be maintained in the blackboard situation due to the co-presence of the material slices, but not, for example, in a PowerPoint presentation. More information related to this topic can be found in the discussion of media ‘depiction’ in Bateman et al. (2017: 126–128), and there are also relationships to be drawn with the ‘sites of engagement’ of mediated discourse analysis (cf. Jones 2005), although slices are present in all communicative situations.

This style of characterization can naturally be extended to a host of further more or less complex multimodal communication situations; in all cases, maintain-

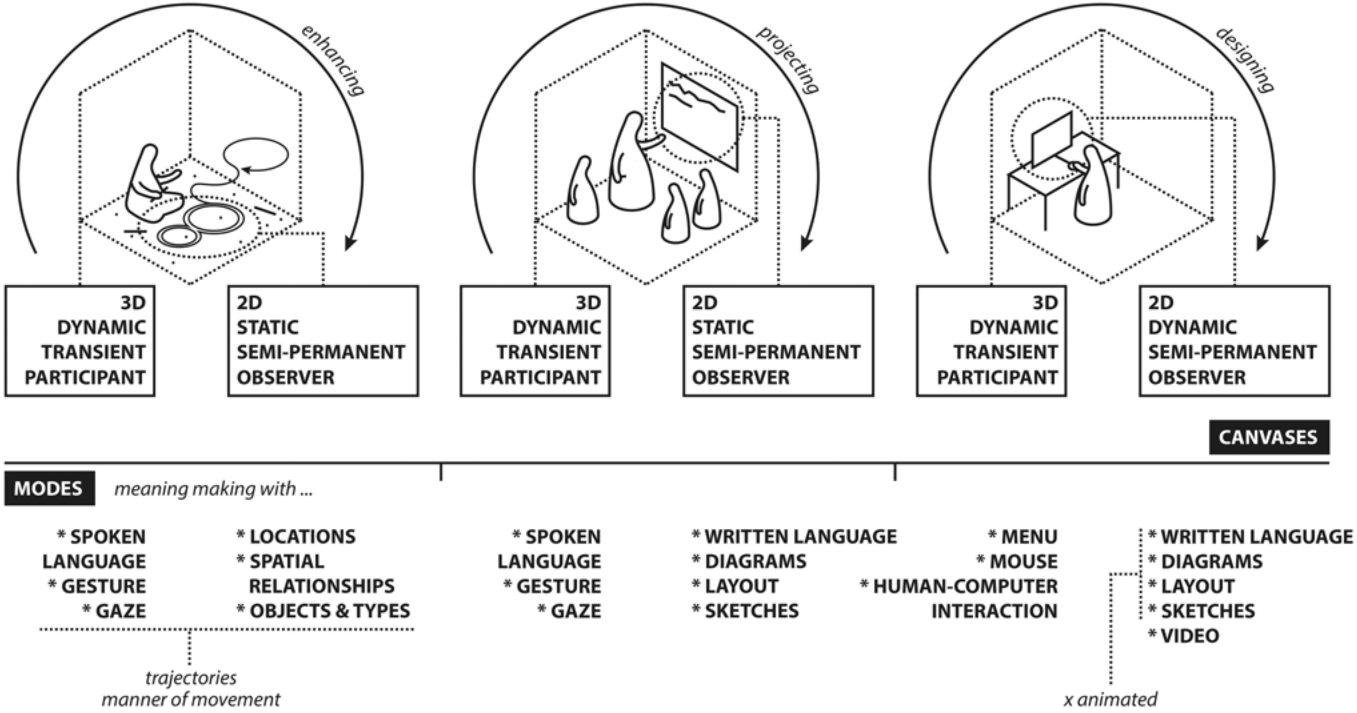


Fig. 4: The three example communicative situations of sand stories, using a blackboard and creating a PowerPoint presentation divided according to material slices (canvases), relations between slices, and potentially afforded semiotic modes. Visualization: Jana Pflaeging.

ing links of similarity and contrast as sketched here supports detailed analysis. The kinds of inter-connections discussed in this section are, therefore, far from sporadic or gratuitous. They arise necessarily (from an appropriate semiotic perspective) as consequences of the similarities and differences in materiality deployed. Building on the extended notion of materiality set out in this chapter thus naturally brings out continuities across situations that previously might have been segmented quite differently. In short, regardless of whether one's internal language of description talks of semiotic modes, semiotic resources, or semiotic systems, anchoring the discussion in materiality may well guide research more effectively.

5 Conclusions and Outlook

In this chapter, it has been suggested that a detailed and semiotically-informed account of materiality can be made to serve as an LCT-style external language of description for the field of multimodality. To show this at work, several semiotically relevant dimensions of materiality were introduced and their use for guiding empirical analysis illustrated. The basic tenet of such an approach is that attending to the fine articulation of the materiality of any object of study offers a robust initial step for subsequent analysis, clarifying just which kinds of material distinctions are available for any semiotic modes assumed to be operating and setting out material boundaries that those semiotic modes may need to work to offset. Furthermore, characterizing the material of any communicative situation in this way establishes a method by which any data gained may be organized for subsequent analysis without premature, and potentially circular, interpretation in terms of those very theoretical categories that are targets of investigation.

Although it is something of a commonplace that accounts of multimodality need to engage seriously with the materiality of multimodal phenomena, systematic treatments of that materiality when shaped and construed for semiotic purposes remain sketchy. There is much discussion of the affordances of such materialities as this is assumed to give strong indications concerning the kinds of meanings that modes using those materialities can deploy but, in many of these discussions, this leads to a certain tension. On the one hand, it is evident that materialities do bring constraints to bear but, on the other hand, semiotic modes appear in many cases quite able to overcome those limitations. For Kress (2010), for example, a consideration of affordances gives rise to two 'logics' — that of space and that of time. But this conflates material properties (e.g., whether an inscription in a material can move or not, is accessible to the senses for inspection, etc.) with semiotic uses of those properties (e.g., the ‘logic’ of narrativization, etc.), thereby weakening potential operationalizations.

Articulating a finer, independent characterization of materiality prior to interpretation of this kind is then far more than an optional extra: without appropriate access to the materiality of objects under investigation, the empirical enterprise as such is weakened. To move on, more robust empirical methodologies capable of securing access to rich and multiple structured multisensorial ensembles for the purposes of analysis must be developed. The dimensions of materiality set out in this chapter aim to provide further critical steps in this direction.

Bibliography

Barker, Jennifer M. 2009. *The Tactile Eye: Touch and the Cinematic Experience*. Berkeley, Los Angeles and London: University of California Press.

Bateman, John A. 2019a. Afterword: Legitimating Multimodality. In J. Wildfeuer, J. Pflaeging, J. Bateman, O. Seizov & C. Tseng (eds.), *Multimodality: Disciplinary Thoughts and the Challenge of Diversity*, 297–321. Berlin: De Gruyter Mouton.

Bateman, John A. 2019b. Multimodality and Materiality: The Interplay of Textuality and Texturality in the Aesthetics of Film. *Poetics Today* 40(2). 235–268. https://doi.org10.1215/03335372-7298536.

Bateman, John A., J. Wildfeuer & T. Hiippala. 2017. *Multimodality – Foundations, Research and Analysis. A Problem-Oriented Introduction*. Berlin: De Gruyter Mouton.

Bernstein, Basil. 2000. *Pedagogy, Symbolic Control and Identity: Theory, Research, Critique*. Oxford: Rowman & Littlefield 2nd edn.

Bezemer, Jeff & G. Kress. 2014. Touch: A Resource for Making Meaning. *Australian Journal of Language and Literacy* 37(2). 77–85.

Björkvall, Anders & A.-M. Karlsson. 2011. The Materiality of Discourses and the Semiotics of Materials: A Social Perspective on the Meaning Potentials of Written Texts and Furniture. *Semiotica* 187(1/4). 141–165.

Bucher, Hans-Jürgen. 2011. Multimodales Verstehen oder Rezeption als Interaktion. Theoretische und empirische Grundlagen einer systematischen Analyse der Multimodalität. In H.-J. Diekmannshenke, M. Klemm & H. Stöckl (eds.), *Bildlinguistik. Theorien – Methoden – Fallbeispiele*, 123–156. Berlin: Erich Schmidt.

Carroll, Noël. 1996. *Theorizing the Moving Image* chap. IV. Defining the Moving Image, 49–74. Cambridge: Cambridge University Press.

Drucker, Johanna. 2013. Performative Materiality and Theoretical Approaches to Interface. *Digital Humanities Quarterly* 7(1). http://www.digitalhumanities.org/dhq/vol/7/1/000143/000143.html (last accessed: 1 September 2021).

Elleström, Lars. 2014. Material and Mental Representation: Peirce Adapted to the Study of Media and Arts. *The American Journal of Semiotics* 30(1–2). 83–138. https://doi.org/10.5840/ajs2014301/24.

Engle, Randi A. 1998. Not Channels but Composite Signals: Speech, Gesture, Diagrams and Object Demonstrations are Integrated in Multimodal Explanations. In M. A. Gernsbacher &

S. J. Derry (eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, 321–326. Mahwah, NJ: Erlbaum.

Forceville, Charles J. 1999. Educating the Eye? Kress and Van Leeuwen's *Reading Images: The Grammar of Visual Design* (1996). *Language and Literature* 8(2). 163–178. https://doi.org/10.1177/096394709900800204.

Fricke, Ellen. 2013. Towards a Unified Grammar of Gesture and Speech: A Multimodal Approach. In C. Müller, A. Cienki, E. Fricke, S. Ladewig, D. McNeill & S. Tessendorf (eds.), *Body – Language – Communication / Körper – Sprache – Kommunikation* (Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science (HSK) 38/1), 733–754. Berlin and New York: De Gruyter Mouton.

Gallese, Vittorio & M. Guerra. 2020. *The Empathic Screen: Cinema and Neuroscience*. Oxford: Oxford University Press.

Gottdiener, Mark. 1995. *Postmodern Semiotics: Material Culture and the Forms of Postmodern Life*. Cambridge, MA and Oxford: Blackwell.

Green, Jennifer D. 2014. *Drawn from the Ground: Sound, Sign and Inscription in Central Australian Sand Stories*. Cambridge, MA: Cambridge University Press.

Hayles, N. Katherine. 2003. Translating Media: Why We Should Rethink Textuality. *The Yale Journal of Criticism* 16(2). 263–290.

Iedema, Rick. 2007. On the Multi-Modality, Materiality and Contingency of Organizational Discourse. *Organization Studies* 28(06). 931–946.

Jewitt, Carey. 2018. Towards a Multimodal Social Semiotic Agenda for Touch. In S. Zhao, E. Djonov, A. Björkvall & M. Boeriis (eds.), *Advancing Multimodal and Critical Discourse Studies. Interdisciplinary Research Inspired by Theo van Leeuwen's Social Semiotics*, 79–93. London: Routledge.

Jewitt, Carey, S. Price, K. L. Mackley, N. Yiannoutsou & D. Atkinson. 2020. *Interdisciplinary Insights for Digital Touch Communication* (Human-Computer Interaction Series). Cham, Switzerland: Springer.

Johnson, Lucy. 2018. Contending with Multimodality as a (Material) Process. *Journal of Multimodal Rhetorics* 2(1). 13–27.

Jones, Rodney H. 2005. Sites of Engagement as Sites of Attention: Time, Space and Culture in Electronic Discourse. In S. Norris & R. H. Jones (eds.), *Discourse in Action: Introducing Mediated Discourse Analysis*, 141–154. Abdingdon and New York: Routledge.

Kang, Seokmin, B. Tversky & J. B. Black. 2015. Coordinating Gesture, Word, and Diagram: Explanations for Experts and Novices. *Spatial Cognition & Computation* 15. 1–26.

Kress, Gunther. 2010. *Multimodality: A Social Semiotic Approach to Contemporary Communication*. London: Routledge.

Kress, Gunther. 2012. Researching in Conditions of Provisionality: Reflecting on the PhD in the Digital and Multimodal Era. In R. Andrews, E. Borg, S. B. Davis, M. Domingo & J. England (eds.), *The SAGE Handbook of Digital Dissertations and Theses*, 245–258. London: Sage. https://doi.org/10.4135/9781446201039.n15.

Kress, Gunther & T. van Leeuwen. 2001. *Multimodal Discourse: The Modes and Media of Contemporary Communication*. London: Arnold.

Kress, Gunther & T. van Leeuwen. 2006 [1996]. *Reading Images: The Grammar of Visual Design*. London and New York: Routledge.

Latour, Bruno. 2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

Ledin, Per & D. Machin. 2019. Doing Critical Discourse Studies with Multimodality: From Metafunctions to Materiality. *Critical Discourse Studies* 16(5). 497–513. https://doi.org/10.1080/17405904.2018.1468789.

Maton, Karl. 2014. *Knowledge and Knowers: Towards a Realist Sociology of Education*. London and New York: Routledge.

Maton, Karl. 2016. Legitimation Code Theory: Building Knowledge about Knowledge-Building. In K. Maton, S. Hood & S. Shay (eds.), *Knowledge-Building: Educational Studies in Legitimation Code Theory*, 1–24. Abingdon and New York: Routledge.

Maton, Karl & R. T.-H. Chen. 2016. LCT in Qualitative Research: Creating a Translation Device for Studying Constructivist Pedagogy. In K. Maton, S. Hood & S. Shay (eds.), *Knowledge-Building: Educational Studies in Legitimation Code Theory*, 27–48. Abingdon and New York: Routledge.

Mersch, Dieter. 2002. *Was sich zeigt. Materialität, Präsenz, Ereignis*. München: Fink.

Mittelberg, Irene. 2017. Experiencing and Construing Spatial Artifacts from Within: Simulated Artifact Immersion as a Multimodal Viewpoint Strategy. *Cognitive Linguistics* 28(3). 381–415.

Mukerji, Chandra. 2015. The Material Turn. In R. Scott & S. Kosslyn (eds.), *Emerging Trends in the Social and Behavioral Science: An Interdisciplinary, Searchable, and Linkable Resource*, Chicester: John Wiley & Sons. https://doi.org10.1002/9781118900772.

Murphey, Keith M. 2005. Collaborative Imagining: The Interactive Use of Gestures, Talk, and Graphic Representation in Architectural Practice. *Semiotica* 156(1/4). 113–145.

Norris, Sigrid. 2009. Modal Density and Modal Configurations: Multimodal Actions. In C. Jewitt (ed.), *The Routledge Handbook of Multimodal Analysis*, 78–90. London: Routledge.

O'Halloran, Kay L. 2009. Historical Changes in the Semiotic Landscape: From Calculation to Computation. In C. Jewitt (ed.), *The Routledge Handbook of Multimodal Analysis*, 98–113. London: Routledge.

Orlikowski, Wanda J. 2007. Sociomaterial Practices: Exploring Technology at Work. *Organization Studies* 28(9). 1435–1448.

Pirini, Jesse. 2016. Intersubjectivity and Materiality: A Multimodal Perspective. *Multimodal Communication* 5(1). 1–14.

Rosenberg, Terry E. 2008. New Beginnings and Monstrous Births: Notes Towards an Appreciation of Ideational Drawing. In S. Garner (ed.), *Writing on Drawing: Essays on Drawing Practice and Research*, 109–124. Bristol: Intellect Books.

Roth, Wolff-Michael. 2000. From Gesture to Scientific Language. *Journal of Pragmatics* 32. 1683–1714.

Schüller, Daniel, C. Beecks, M. Hassani, J. Hinnell, B. Brenger, T. Seidl & I. Mittelberg. 2017. Automated Pattern Analysis in Gesture Research: Similarity Measuring in 3D Motion Capture Models of Communicative Action. *Digital Humanities Quarterly* 11(2). http://www.digitalhumanities.org/dhq/vol/11/2/000309/000309.html (last accessed: 1 September 2021).

Scollon, Ron. 2001. *Mediated Discourse: The Nexus of Practice*. London: Routledge.

Sobchack, Vivian. 1992. *The Address of the Eye: A Phenomenology of Film Experience*. Princeton: Princeton University Press.

Streeck, Jürgen. 2013. Interaction and the Living Body. *Journal of Pragmatics* 46(1). 69–90.

Tseng, Chiao-I. 2017. Film Space as Theatrical Performing Space: A Multimodal Discourse Approach to Transmedial Analysis. In M. G. Sindoni, J. Wildfeuer & K. L. O'Halloran (eds.), *Mapping Multimodal Performance Studies*, 127–153. London and New York: Routledge.

Tversky, Barbara & A. Kessell. 2021, this volume. Thinking in Action. Reprint. In J. Pflaeging, J. Wildfeuer & J. A. Bateman (eds.), *Empirical Multimodality Research. Methods, Evaluations, Implications*, 91–108. De Gruyter Mouton. Reprint from Pragmatics & Cognition 22.2, 206–223.

van Leeuwen, Theo. 1999. *Speech, Music, Sound*. London: MacMillan.

van Leeuwen, Theo. 2005. *Introducing Social Semiotics*. London: Routledge.

van Leeuwen, Theo. 2009. Parametric Systems: The Case of Voice Quality. In C. Jewitt (ed.), *The Routledge Handbook of Multimodal Analysis*, 68–77. London: Routledge.

John A. Bateman and Tuomo Hiippala

From Data to Patterns

On the Role of Models in Empirical Multimodality Research

Abstract: Most approaches to empirical multimodality research adopt the assumption that communicative situations will give rise to distinctive patterns of phenomena which may be captured by formulating appropriate models and frameworks. While this assumption may be true in principle, detecting such patterns is far from straightforward. Informative patterns are likely to be distributed both within and across semiotic modes, be spread across various levels of abstraction, and exhibit intrinsic internal variation. The highly elusive nature of patterns arising amidst such complexity emphasizes the need for increased attention to *modeling*, that is, characterizing more closely the relationship between the theoretical constructs defined and the phenomena under analysis. In this chapter, we relate the practice of modeling to semiotic principles and introduce one of the current methods for drawing patterns from data that is gaining increasing traction in empirical studies across the board.

Keywords: modeling, Peirce, iconicity, statistical methods, multimodality

1 Introduction

Research, particularly empirical research, is about finding patterns that help us understand and explain underlying processes or mechanisms. One might want to show, for example, how certain grammatical patterns correlate with variations in power and other social configurations, or how the spatial distribution of information across a news website correlates with news values, or how particular body movements, gestures, and prosodic patterns may be considered indicative of particular communicative intentions or situations. Van Leeuwen (2005a: 4-8) suggests that one might begin research of this kind by making observations and progressively organizing those observations into systematic classifications. Such classifications are often developed in pilot studies and then extended as necessary for any particular data under investigation. Van Leeuwen claims further that such processes of extension generally involve adding or refining categories, but will rarely require revision of what has been done before (van Leeuwen 2005a: 14). It is unclear, however, why such classificatory infallibility should occur. More likely, at least *prima facie*, would be that either interpretation is being distorted by confirmation bias or the categories employed are sufficiently weak that they may be 'bent'

https://doi.org/10.1515/9783110725001-003

to fit data as necessary. Neither situation offers a particularly robust foundation for further research.

At some point, therefore, one must ask whether any categories or classifications proposed are (a) sufficient in scope and (b) actually revealing of patterns observable in the data. The word 'revealing' is intended here in a very specific sense: that is, can the patterns identified be used *predictively* to reason about the phenomena at issue? This requires us not to be satisfied simply with descriptions of data — one can, after all, always label phenomena; we need also to be able to make effective predictions about any patterns observed. This moves us from observations of data to hypotheses of explanatory conditions that can tell us whether specified changes in conditions are likely to give rise to observable changes in patterns, and *vice versa*. Knowing this with some reliability is a prerequisite for designing interventions of any kind and for generating further research questions.

Interest in asking and answering questions of this kind is already widespread in multimodality research and constitutes a driving force behind a corresponding growth in work that explicitly aims at being 'empirical' in orientation. This development is not, however, proceeding as rapidly as it arguably should given the goal of reliably relating patterns to conditions. One reason for this is the extreme breadth of the study of multimodality — since the primary goal is to explain how material of *any* kind may be used to communicate, or 'signify', the domain of study turns out largely commensurate with that of semiotics as a whole (cf. Nöth 1995). This naturally involves research from many disciplines spanning a rich variety of focuses, goals, and methods (cf. Jewitt 2014).

A significant feature distinguishing multimodality research from semiotics more broadly, however, is precisely its orientation to data. As a form of *applied* semiotics, multimodality research of all kinds seeks to further its concerns by paying close attention to observed instances of (multimodal) communication in action.[1] Nevertheless, patterns are still generally observed and described in the manner familiar to each tradition and this results in relatively little consensus concerning how the aims of multimodality research can best be achieved over all. Some approaches extend from forms of experimental methods, as widely practised in psychology, whereas others draw on distribution-based data, which itself may range from single case studies, to small sets of illustrative examples, to large-scale 'corpus' analyses. Others again are more sociocultural in orientation and rely on broadly textual descriptions and explanations of any phenomena studied. Whereas

[1] This does not preclude more theoretical, or philosophical, orientations within multimodality research; such work would, however, then necessarily contribute to semiotics more broadly as well.

this diversity is, in principle, to be welcomed, the methodological differences in play have also led to boundaries being drawn along more traditional disciplinary lines, invoking distinctions such as empirical/hermeneutic, qualitative/quantitative, social/cognitive, and so on. Such divisions can impede progress by preventing distinct approaches and results from productively engaging with one another.

Despite some prominent calls for the application of pluralities of approaches (e.g., van Leeuwen 2005b), multimodality researchers still far too often adopt the theories and methods of the ‘home discipline’ within which they received their primary training. Such methods may not, however, necessarily offer sufficient leverage for dealing with the phenomena they target since those phenomena may well lie outside the home discipline’s object of study. Our concern in this chapter will therefore be to re-emphasize that, just as multimodality benefits from its breadth of concerns, it can equally benefit from a broader application of a diversity of methods *across traditionally inherited disciplinary boundaries*. To support this development, we will consider the role of *modeling* as a central contribution to effective empirical multimodality research. We begin by framing the notion of models against the backdrop of Peircean semiotics to characterize more precisely the kinds of insights and knowledge that models can help produce. We then outline some illustrative procedures for empirical research that implement productive dependency relations between theory, models, and data, casting this from a *semiotic* perspective throughout.

2 Looking for Explanations: A Peircean View of Models

Whatever kind of investigation of multimodality one is concerned with, one is generally working towards explanations of some phenomena under observation. This means that one has identified an area of interest where some regularities appear to be at work, and one attempts to describe those regularities in a way that allows predictions to be made about their behavior. Such descriptions clearly need to exhibit certain properties of their own. Most importantly, they need to be able to serve as *tools for thinking* about their objects of concern — that is, the descriptions need to find echoes in what one is describing so that they can drive new ways of understanding what is going on. We will refer to such descriptions as *models*. A model of some phenomena is thus a systematically produced simplification of the full complexity of those phenomena that can ‘stand in’ for that full complexity for purposes of reasoning and explanation. Put differently, a model is a focusing device which deconstructs the full complexity by picking out entities, relationships, processes, and mechanisms that appear to be usefully ‘congruent’ in some respect with what is being investigated.

Such focusing devices vary considerably in form: indeed, taking a multimodal perspective on this, a model could be realized equally as, for example, physical objects or sets of mathematical equations. In much multimodality research hitherto, models have been couched in textually constructed configurations of interrelated theoretical constructs, which together support informative ‘readings’ of multimodal phenomena. Such models also vary considerably in scope: some seek to describe the generic foundations underlying all forms of multimodal communication (e.g., Kress & van Leeuwen 2001; Norris 2016; Bateman et al. 2017), while others target specific forms of expression, such as text and image (Martinec & Salway 2005) or spoken language and gesture (Kendon 2004; Fricke 2012). Given this diversity, it is natural that there is continuing discussion concerning how the notion of ‘model’ should best be defined.

Our mentioning of the notion of congruence above restricts the scope of potential solutions, however. Models must be seen as specifically structured descriptions which *by virtue of their structure* allow new knowledge to be gained about their target. This makes the task of definition significantly easier. In fact, the question: ‘What are models, no matter from which discipline?’ leads inevitably back to semiotics. Models pose an inherently semiotic question precisely because they are (possibly abstract) artifacts that, for their interpreters, refer in certain respects and through specific properties to something else — which is simply to reiterate Peirce’s general definition of a sign (e.g., Peirce 1998 [1893–1913]: 478). Models are therefore signs par excellence and cannot be anything but semiotic. Moreover, models meet the requirements of one particular way of being a sign defined by Peirce: that of ‘iconicity’.

Signs functioning iconically allow statements to be made about some object by virtue of properties of the icon itself. If a sign does not perform this function (or to the extent that it does not perform this function), then it is not functioning as an icon. Iconicity exists, therefore, precisely when the ‘intrinsic’ properties of the sign vehicle itself may be used to indicate properties of the sign’s object — as when a picture of a tree might be used to tell an interpreter that the leaves are green or that the branches are twice as long as the trunk. Although iconicity is often thought of in terms of pictorial representations and visuals, Peirce’s definition is in fact far broader: any use of one entity (the representamen or ‘sign vehicle’) to give us information about another (the object) by virtue of properties inherent to the first entity (i.e., the representamen) meets Peirce’s definition. In this chapter, we will be referring to iconicity in this most general sense throughout. Connections between models and Peirce’s definitions of iconicity have been drawn by several commentators — Kralemann & Lattmann (2013), for example, consider models within the digital humanities, while Ambrosio (2014) addresses models in science more generally. The power of iconicity for conceptualizing models is made particularly clear by Hookway:

> The key of iconicity is not perceived resemblance between the sign and what it signifies but rather the possibility of making new discoveries about the object of a sign through observing features of the sign itself. Thus a mathematical model of a physical system is an iconic representation because its use provides new information about the physical system. This is the distinctive feature and value of iconic representation: a sign resembles its object if, and only if, study of the sign can yield new information about the object. (Hookway 2000: 102)

This perspective needs to be taken instead of the far weaker (and long critiqued) notions of iconicity as 'likeness'.

Different types of models may be more or less well suited to particular kinds of methodological approaches, or remain open to selecting those methods that best match the research questions asked. The most open, or least restrictive, form of model may be characterized as discursive: this means that the analyst produces a description of the intended theoretical entities and those entities' interrelationships in the form of a piece of (usually) written verbal language (e.g., Kress & van Leeuwen 2006 [1996]; Serafini & Reid 2019). The purpose of a description of this kind is to provide construals for any particular phenomena under discussion in terms that are given by the model. The descriptions constituting such models are qualitative in the sense that the categories and relations used make qualitative distinctions between situations or patterns. Examples of qualitative distinctions are given/new, elaboration/circumstance, hard news/editorial, and so on. Qualitative categories therefore group together collections of phenomena and make the statement that, for the purposes of the category, the members of those collections are to be treated as equivalent or similar to each other. These descriptions can also offer (textual) explanations by stating that particular configurations may give rise to, or be caused by, other configurations and so constitute an essential component of generalizing beyond particular cases or instances. It is, nevertheless, up to the analyst to specify how the constructs of the model relate to observable data, and this can be done with greater or lesser precision (cf. the discussion in Bateman, this volume).

The languages in which models are expressed may also have differing properties. Whereas a purely textually expressed model relies upon 'textual' logic, i.e., commonsense reasoning, to make its statements, models expressed in more formalised languages may support more specific patterns of inference. For example, a model expressed in the predicate calculus supports the application of principles of reasoning from formal logic. This can make it more straightforward to evaluate whether the model has gaps or makes nonsensical predictions, even before being confronted with data. One might also have purely mathematical models where the relationships between categories are expressed by mathematical relations and structures of various kinds. The more 'generative' a model becomes in its own right, the more strongly inferences made following the logic of the model are simultaneously hypotheses and conjectures concerning expected patterns in the data. This moves a characterization from being a description, that is, a system of labels for phenomena, to a generative theory that predicts how phenomena will pattern, even if those patterns have not yet been observed. As such, the resulting blend of model and data may be described as *meaning-generating* because patterns in data are bound together with the model's descriptions and explanations of those patterns.

The nature of the relationship of models to the data that those models are intended to characterize is therefore an important aspect of methodology. The more closely that a theoretical language and its models are tied to data, the more it is appropriate to consider the research as *empirical*. Here the blended nature of data and descriptions moves centre stage: the model will generate hypotheses concerning the data and, if the data is found to be incompatible with those hypotheses, then the model needs to adapt. This is the usual notion of the 'empirical cycle', sometimes inaccurately associated with purely quantitative research but in fact applicable to all investigations that consider data external to the theory itself. In the empirical cycle, hypotheses or conjectures are generated from the model, these statements are mapped to patterns in the data, and the data is examined to see if those patterns occur; if they do, then the hypothesis is supported (or at least not rejected) and if they do not, then evidently there is a problem. That problem may lie in the model used or in the way of relating the model to the data. Addressing the problem is almost certainly guaranteed to improve understanding of the phenomena at issue. It is worth emphasizing, therefore, that it is (almost) never too soon to start considering the empirical cycle as a way of refining explanations and theories. This can start with extremely broad qualitative categories that are only successively made more precise — the crucial step for empirical research is to begin.

A final aspect playing an important role for the use of modeling that we will consider is the question of *evaluating* the extent to which a model operates predictively: that is, one wants to achieve situations in which the conjectures or hypotheses generated by a model are *reliable* with respect to the corresponding patterns or regularities found in the data. Whereas an empirical model is a model that is explicitly related to data, a *good* empirical model is one that exhibits a high degree of predictivity with respect to the data. One might make all sorts of statements about how some body of multimodal data is predicted to pattern, but if the connection between those patterns and the model's hypotheses is insufficiently reliable, then the empirical cycle cannot be used to progressively refine results. We need, therefore, to consider questions of reliability explicitly as well: in short, *how well* is a model achieving its function as an icon of generating new knowledge?

3 Empirical Procedures: From Theory to Models to Data and Back

As indicated above, most empirical approaches to multimodality adopt the assumption that communicative situations may be characterized by distinctive patterns that leave 'measurable' traces; these patterns can then be captured by developing models and frameworks appropriate for their analysis. While this assumption may be true in principle, detecting patterns and validating their presence is challenging because patterns are likely to be distributed both within and across semiotic modes, and in any case exhibit natural variation. The elusive nature of patterns emphasizes the need to pay increased attention to the process of modeling: this involves not only making the relationship between the theoretical constructs defined by some model and the phenomena under analysis explicit, but also actually using those models to derive patterns and evaluating the reliability of the results. For empirical work, therefore, models should function as tools for probing which descriptions and which interdependencies between concepts can best characterize some body of data. In short, we need to find models that: (a) can 'stand in' reliably for actually observed patterns of regularity in any data by predicting them (iconicity) and (b) can render those patterns intelligible by offering explanatory re-descriptions of the predictions in terms of the abstract concepts and mechanisms defined in the models and corresponding theories. This means that models relate both to theory and to data, mediating between them.

If we think of what predictions might be, then most abstractly we want to be able to relate patterns in any data to conditions and circumstances such that (specifiable) changes in conditions lead to, or correlate with, particular changes in observable patterns. What we then consider a 'good' model will be a model that reliably predicts how circumstances and patterns correlate. As models improve, then their predictions should become finer, i.e., more discriminating with respect to both conditions and outcomes, and more accurate. Modeling of this kind is necessarily a process of continual refinement. As set out by Peirce and many subsequent philosophers of science, any model is simply a current 'best estimate'. Alternatively, in terms of an adage attributed to Box (1979): "All models are wrong but some are useful". Our goal must then be to find methodologies which can support the development of classes of models that offer predictions of patterns in data and which thereby themselves constitute more abstract characterizations of those patterns. Critically, such characterizations have little to do with notions of ‘objective truth’ or ‘certainty’, as sometimes alleged concerning empirical methods – such nods to positivism are not compatible with the essentially Peircean position we maintain throughout.

As a starting point, then, let us assume that we have some more or less complex qualitative theoretical framework that provides a way of interpreting phenomena in some selected multimodal situation or artifact. This theory will generally include a range of terms linked to intended interpretations and back to instances of data: that is, the theory allows observations to be ‘read’ in terms of the theory. We propose that any qualitative and discursively defined theory can be used to generate partial models that can be interrogated empirically. Such partial models may be constructed by increasing the specificity of the general model in particular areas of interest so that it becomes possible to formulate hypotheses in terms of predictions of patterns in data. Performing this step will often already reveal aspects of the original theory and its models that require more work, refinement or revision; this is one additional motivation for carrying out this step in the first place.

Several basic methodological tasks can be defined to support the development of such partial models. We sketch these first in terms of their relations to theories and models and then, in a second cycle, illustrate some established methods by which they can be performed in practice. Although these stages and methods for conducting empirical work on a theory are well known in empirically-oriented disciplines, they are still too often neglected in several prominent branches of multimodality research. As suggested above, this is sometimes due to the disciplinary boundaries involved, whereby the relationship between abstract qualitative theories and models for which empirical methods are appropriate is misconstrued as a methodological divide rather than an independent choice. In short, nothing prevents the application of these methods across all forms of multimodality research — which is *not* to say, of course, that all research activities must always use all research methods! The following tasks, then, specifically characterize empirical research:

- First, there needs to be some body of data that is thought relevant for the theory being investigated, that is, the theory makes statements about data of that kind. This is the essential ingredient of any work that considers itself empirical. Note that the data considered can be either naturally occurring, leading to corpus-based research, or specifically manipulated for experiment-based research.
- Second, on the basis of the theory, any data considered needs to be segmentable into units of some kind. These units can be of any scale, up to entire texts and beyond, but the criteria for recognizing units must be made as explicit

and reliable as possible. Put differently, working with the criteria, different analysts should make the same segmentation choices to a degree that is significantly greater than chance (cf, Bateman et al. 2017: 198-204). This helps 'operationalize' the process of reading data through theory.

- Third, qualitative categories or measurements need to be formulated with respect to what the theory suggests should be useful or relevant. These categories or measurements are then applied to each and every unit. This produces a body of data in which each identified unit has received a collection of classifications or measurements describing that unit. As with the second step, these classifications and measurements need to be made as reliable as possible. The result here is often an annotated corpus, in which the annotations capture the technical features of a semiotic mode, or are assumed to lead to the formulation of such technical features on the basis of theory and further analysis.
- And fourth, the actual search for patterns must take place. This can proceed both 'bottom up', by employing methods to see to what extent the annotations allow that data to be grouped into clusters, and 'top down', by adopting or generating further sets of categories from the theory that the theory suggests may serve a *conditioning* role for the technical features or annotations established in the third step. Patterns may be said to occur whenever the categories of this fourth step correlate significantly with the categories of the third step. In other words, collections of technical features (and the units they classify or measure) are grouped into equivalence classes given by the conditioning categories. This captures similarities and differences across the units in the data: the organization of data employing features from this step can 'stand in' as qualitative descriptions of the data with respect to the patterning of their technical features (third step).

The conditioning categories that serve as anchors for patterns may either be 'external' to the data or be forms of further categories derived from the theory for the description of units. External categories label aspects of the production situation of the units in the data and so might include features related to geography, social classes, medical conditions, time periods, positions on pages or screens, and so on. In contrast, internal categories are more related to particular units in the data and might include assumed conditioning factors such as genre, rhetorical relations, boundaries between units, degrees of visual modality, classes of text-image relations, and so on. All such features are similar to the kinds of labels used in content analysis and usually need to be mutually exclusive and exhaustive (cf. Schreier 2012). Whenever correlations can be documented between the conditioning features and the technical features, the conditioning features can also be used predictively for the technical features. In fact, correlations can be sought among any collections of classifying features. Thus one might predict that in a certain genre, particular rhetorical relations or text-image relations might be more frequent, or that within a particular degree of visual modality certain visual technical features, such as hue and brightness, might receive certain ranges of values rather than others, and so on. To the extent that such expectations, or hypotheses, are not met, one is led back to consider revisions of the account. This constitutes an abstract characterization of the empirical cycle.

4 Methods for Finding Meaningful Patterns

As mentioned above, disciplines that are already empirically oriented have a long tradition of developing methods for finding patterns, using those patterns as predictions, and evaluating those predictions' reliability. In this section, we argue for the potential relevance of this tradition by considering a particular class of methods and models that is currently gaining attention in several disciplines related to multimodality research, but which remains rarely applied in multimodality research itself. We show that these models can be applied equally to multimodal empirical research and argue that adopting them more broadly would significantly improve our ability to find meaningful patterns when building theory. Similar arguments can be made for many such approaches from empirically-oriented disciplines.

In established empirical sciences, models are often seen in mathematical terms because this makes it relatively clear how 'predictions' of quite complex patterns can be formulated. They are thus strongly 'generative' in the sense introduced earlier. This may previously have obscured the relevance of these methods for multimodality research, which still, as noted, often constructs its conjectural qualitative models using textual descriptions. There is, however, no need to maintain methodological boundaries between these complementary ways of pursuing meaningful patterns. Instead, our characterization of models as semiotic constructs bridges the gap between discursive multimodality research and empirically-based model building. Theories and models, regardless of the languages in which they are formulated, are inherently semiotic because they 'stand in' for the phenomena under investigation. There is, therefore, no need to artificially restrict the kinds of descriptive languages that we can employ when articulating these theories and models — indeed, to do so is likely to make empirical work simultaneously more difficult and less effective. To proceed, we focus specifically on the fourth task in the list of tasks we set out in the previous section because the others have all received attention elsewhere (e.g., Bateman et al. 2017: Chap. 6; Bateman & Hiippala 2020). The fourth step, searching for patterns, has received far less attention.

4.1 Regression Models

As described above, the search for meaningful patterns can be construed as a continual process of making predictions of distributions of observable properties on the basis of conditions given by theory. For example, we might assume that our theory predicts that a category of 'genre' should make a difference for some technical features deployed in the data, such as, e.g., the quantity of visual material included. Then, becoming more concrete, we might postulate that our data comes from three distinct genres and we check in our annotated multimodal data whether this conditioning feature (genre) actually correlates in some way with the number of photographs present. If our theory suggests that the number of photographs should vary according to genre, then we should see this pattern in the data as well. If we do not find such a pattern, then more refinement or alternative conjectures are necessary. This is consequently a deliberately 'theory-directed' approach to exploration.

In addition to wanting to know whether some conditioning factor makes a difference, we are often interested in knowing more about that difference. For example, rather than simply knowing that a change related to genre will make a difference, we want to say something about *just what kind of difference* follows from that change — i.e., we might want to ascertain whether certain kinds of technical features increase in frequency and others decrease, and so on. We see this as a stronger notion of prediction that more directly relates conditioning factors to observable consequences. The particular class of models supporting this kind of exploration that we have selected to address here is that of so-called *generalized linear models*. These models are now receiving increased application in a broad variety of disciplines (Field et al. 2012; Winter 2020) but are still rarely seen in multimodality research. The basis of this kind of modeling is deceptively simple and involves a generalization of the mathematical notion of 'regression'. Regression is when one looks to see if some set of values correlates with some other set of values. Correlations and how to calculate them are described in all introductory textbooks on statistics, but since our concern here is to show how this technique naturally extends into very diverse areas of model-building that are not yet so widely known in multimodality research, we set out the basic tenets briefly before proceeding.

Correlation is concerned with describing *co-variation*: that is, if one thing varies in a particular way, something else varies correspondingly. Two sets of values are then said to correlate positively if increasing values in one set (for example, brightness of a photograph) go together with increases in values in the other set (for example, duration of looking at the photograph). Alternatively, two sets correlate negatively if an increase in one goes together with a *decrease* in the other. Finally, two sets do not correlate if one cannot make any statement about relative increases or decreases. There are, nowadays, many computational tools that make the calculation of correlations straightforward and the general idea can be readily extended to consider any number of sets of potentially co-varying data. Considered more semiotically, finding correlations by regression works in essence by deriving a *re-description* of the (sets of) data in simpler terms. If, for example, we take our brightness values of images and the lengths of time spent looking at each image, we can represent this on a graph where one axis reports the brightness values and the other axis reports the duration lengths. Each point in the graph then captures the pairing of brightness and duration for some given image; such a collection of points is shown on the left-hand side of Figure 1. A far simpler re-labelling of all these points is then given by calculating a *regression line* as illustrated on the right-hand side of Figure 1.

A 'linear' regression line of this kind describes the relation between the two sets of data by finding the line that passes 'best', i.e., closest, to all the data points. In other words, this line collectively minimizes its distance to all the individual points on the graph — and so can be uniquely calculated. The resulting line is then a *re-description* of the data in that it can stand-in for the dataset as a whole: for each brightness value (the position on the horizontal axis, or x-axis), it 'predicts' a duration length (the position on the vertical axis, or y-axis, of the graph according to the line). This can naturally succeed better with some data than with others. A very good 'fit' would be when all the points lie on or close to the line and so the predictions will be close to the actual values; a bad fit would be when the points all lie a considerable distance away from the line.

The regression line is, moreover, formally a *simpler* re-description of the data because straight lines can be uniquely characterized by saying where they cross the vertical axis in a graph (called the 'intercept') and how steep they are (called their 'slope'). The entire set of values in our dataset and their combinations has consequently been reduced to just two numbers: the intercept and the slope. In the case shown in the figure, therefore, we can see that the regression line crosses the vertical axis (i.e., the place where the horizontal value is zero) at 14.53. The slope of the line was calculated to be 0.48, which means that for every 1 unit of change along the horizontal (x) axis, there is an increase of 0.48 units on the vertical (y) axis. Concretely, this means that if we move from the horizontal position $x = 0$ to the horizontal position $x = 10$, the value of y will increases from the intercept (14.53) to 10 times the unit slope increase (i.e., $14.53 + 0.48 \times 10$) — which comes

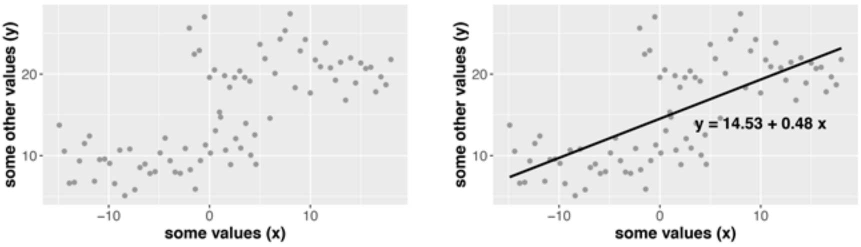


Fig. 1: On the left we see a random collection of points formed by pairs of values: one of each pair on the horizontal axis, the other on the vertical axis; on the right we see the result of calculating a regression line through these points. The equation shown is the mathematical description of the calculated line: given a value of x, the equation gives the corresponding value of y.

out as 19.33 as can readily be verified by looking up the value of y in the graph when x is 10. Finally, in Peircean semiotic terms, the regression line is clearly *iconic*, because knowing about the properties of the line tells us more about the data (see Section 2). In this sense, the regression line is a *model* of the data and the intercept (14.53) and slope (0.48) give the *coefficients* of that model. Using this information, we can predict values and, by virtue of the collected distances of points from the line, measure the accuracy of those predictions: the smaller the collected distances, the closer the points are to the line and so the more accurate the predictions.

Many computational tools are available that allow linear models to be calculated very easily. In this chapter we will use the programming language R throughout (R Core Team 2016); all the graphs shown are also produced directly with R. For the data shown in the figure, therefore, if we store the data in a dataset labelled 'data_pairs' and call the brightness values (the positions horizontally in the graph) 'brightness' and the duration values (the positions vertically on the graph) 'duration',[2] then we can use R's model constructing function 'glm' (generalized linear model) to produce the regression model shown on the right of Figure 1 by writing the following:[3]

```
glm( formula = duration ~ brightness, data = data_pairs )
```

[2] Simple instructions for how to do this are given in Bateman & Hiippala (2020) and the associated website.

[3] Although for this simple example we would normally just use a linear model and the function 'lm', we take 'glm' here to emphasize consistency across all our examples.

This tells R to produce a model in which the brightness values (as indicated on the right-hand side of the formula) are used to predict the respective duration values (as indicated on the left-hand side). The details of the model R produces can be examined by telling R to print out a summary, which in abbreviated form looks approximately as follows:

```
Coefficients:
            Estimate    Std. Error    t value    Pr(>|t|)
(Intercept)  14.5311        0.5025     28.062     < 2e-16 ***
brightness    0.4810        0.0563      7.876    1.52e-11 ***
```

The `Estimate` column gives the coefficients defining the model; here we find the intercept (14.53) and the slope (0.48) just as described above; we return to the other values below. Note that this procedure remains the same for any data that may have been gathered: as long as there are two sets of corresponding values then R can calculate a linear model. The question of *whether it makes sense* to consider those values together remains a theoretical decision — that is, models should always be pursued with theoretical motivations in mind.

Two final points now need to be added, both crucial to model building. First, although the example examined here may seem to be relevant for a very limited range of possible model-building scenarios, i.e., cases where there are two sets of measurements to compare, we will see below how this method generalizes naturally to take in a far broader range of modeling situations, including situations that are qualitative in nature. And second, if the example above were an actual empirical study in which we were interested in seeing if there was a relationship between brightness and length of time spent looking at a photograph, it would never be possible to look at ‘all’ photographs to see if the predicted relation always holds. Empirical work of necessity concerns itself with *samples*, i.e., selections of data, and so it is an important part of modeling to ascertain how likely it is that some calculated model will generalize beyond the immediate cases examined.

The fact that we cannot examine all the cases of interest when constructing a model means in particular that it becomes necessary to make some assumptions concerning the data that we do *not* have. We can always calculate an exact, unique regression line for any data that we have gathered, but this can of necessity only be an approximation of the data that we do not have. Linear models then operate by making some quite specific assumptions about the unseen data and, on the basis of those assumptions, *estimating* both the intercept and slopes associated with any contributing factors. This is why the column in the summary of the model given above was called `Estimate` — that is, they are estimates not because the calculation of the regression line is approximate — for any given set of data points there will be one and only one unique regression line as a result — but because, in empirical research, the data we can collect is always incomplete. There may well then be better regression lines that characterize the data more precisely when we have information about how that data patterns more broadly.

The other values in the summary table above then offer a precise indication of how confident we can be that the estimates offered will extend to further, unseen data. This is, therefore, a very powerful methodological step. It is made possible by considering just how varied the data that we do have is. If values in the observed data are wildly varying, then it is more likely that further data we encounter will also be wildly varying and so the estimates calculated may be considerably off the mark. Conversely, if the values we have appear less variable, then our confidence in the estimates increases. The column `Pr(>|t|)` then gives a probability calculation for how likely it would be that the estimates for intercept and slope have no reliable influence on prediction, given the variation already observed in the data. The values in the column include 'e', for exponent, which is R's way of writing very large or very small numbers as multiples of powers of ten: that is, the expression '1.52e-11' means 1.52×10^{-11}, which is a very small value indeed.[4] In the present example, therefore, the probabilities that the estimates have no reliable influence are both extremely low. This means that we can be correspondingly confident that there is indeed a *reliably predictive relationship* between the two variables. The stars in each row of the summary table give an additional graphic indication of the degree of statistical significance involved (e.g., *** for at most 0.001, ** for at most 0.01 and * for at most 0.05). The remaining columns of the table then offer indications of just how much the estimates can be expected to vary, even though they will be broadly reliable.

It is only possible to calculate such probabilities by making assumptions concerning how the data is expected to pattern 'overall'. Linear models assume that the values for the intercept and slopes can be described with respect to what in statistical terms is called the normal distribution. In a normal distribution, the values follow a 'bell curve' frequently illustrated in statistics introductions (Bateman & Hiippala 2020: 4-5). The assumption that the values estimated for the intercept and slopes will be distributed normally makes certain calculations for producing a linear model more straightforward, but at the obvious cost that some particular data we examine may not fit. This will then be an additional source of errors and uncertainty when using the model for prediction. Nevertheless, one good reason for adopting the assumption of a normal distribution is that this is, technically, an assumption of 'least commitment' — that is, if we know

[4] The value $1.52 \times 10^{-11} = 0.0000000000152$.

nothing about the data and how it is expected to pattern, then assuming the normal distribution is a way of expressing this lack of knowledge. Semiotically, it is the modeling assumption that assumes the least and so is a good choice unless we have more information to bring to bear, either from the collection of more data or from theory.

4.2 Generalized Linear Models: From Correlations to Differences

We now return to the point made above that the application of modeling of the kind described above is far broader than might be apparent at first glance. We noted that linear models make certain assumptions concerning the properties of the models constructed, namely the assumption that the model coefficients follow a normal distribution. In the far broader class of *generalized* linear models, these assumptions are relaxed. Many other kinds of statistical distributions can be considered when searching for a best model and this has immediate consequences for the kinds of data that they can be applied to. R allows such models to be calculated automatically, even when many factors are potentially involved. That is, one can postulate the relevance of particular factors for an outcome and then calculate both the best possible model relating those factors and the observed outcomes and the degree to which that best possible model would indeed 'predict' the data.

The factors that can be considered when deriving linear models need not then be restricted to numbers, such as those taken in the example above. As we shall now see, they can equally easily be made to correspond to (i.e., to model) *qualitative* categories as well, such as genres, information statuses, camera angles and distances, degrees of modality, attributions of power, and so on. Therefore, although our discussion above of fitting lines to data points may sound 'quantitative', it is essential to see that what is being described by such models is equally construable as *characterizations of patterns of change and difference*. This kind of model is of extremely broad application for all kinds of research questions and constitutes a significant generalization beyond the simple numerical situation illustrated in Figure 1. This makes generalized linear models a powerful tool for empirical research questions at large (Cohen 1968). Reworking the quotation from Hookway (2000: 102) in Section 2: a mathematical model of a *semiotic* system is an iconic representation, because its use provides new information about that semiotic system. This is precisely the role we now attribute to generalized linear models.

Given some annotated data and a collection of conditioning features, then, the most basic question to be asked is whether the conditioning features actually have any effect. That is, if one looks at units of analysis grouped under one of the conditioning features and units of analysis grouped under another of the conditioning features, do those units exhibit any differences? If one finds that the same kinds of units are quite happy to occur under differing conditioning features, then it is evident that one cannot use those conditioning features as a meaningful characterization of the data *with respect to those technical features used for description*. This does not, of course, rule out the possibility that other descriptions of the data might show themselves to be sensitive to the conditioning factors investigated, but until this is shown, it is wise to be cautious.

To illustrate, we will use an example from a richly annotated multimodal corpus of 1000 primary school science diagrams (Hiippala et al. 2020). This corpus describes diagrams at several levels of abstraction, including features of their technical composition, visual genre, rhetorical organization, rhetorical complexity (labelled 'RST entropy', which is calculated on the basis of the rhetorical relations found to hold between the elements of each diagram and their diversity), and more, and so allows many issues of multimodal import to be explored. Here we show how the formulation of a generalized linear model can be used to address the research question of whether visual genres differ with respect to their rhetorical complexity. This might then form part of a broader discussion concerning the realization of visual genre and how it might be recognized in different contexts.

Considering the data, the visual genre labels are clearly qualitative categories, while the rhetorical complexity value is a numerical measurement. This means that for each observation in the qualitative visual genre category, we have a set of complexity values. Our research question can therefore be made more precise as an assessment of whether, and to what extent, visual genres may be reliably associated with rhetorical complexity. Taking just two of the visual genres for the purposes of discussion, 'Network' and 'Table', we can obtain an overview of the data by plotting the complexity values for each diagram classified according to each genre as shown in the left-hand graph in Figure 2. Each dot in this graph shows the value of rhetorical complexity for a specific diagram. The dots are grouped according to visual genre and so we have two vertical groupings, each grouping containing the values for the corresponding diagram type.

Now, from this graph alone, it is unclear whether a difference between the visual genres has been found or not. This is a typical situation in empirical research: just obtaining a collection of results and visualizing them may not immediately reveal patterns. A better impression can be gained by adding to the graph the results of basic descriptive statistics, as shown in the middle graph in Figure 2. This box plot visualization adds diagrammatic indications of median value and the

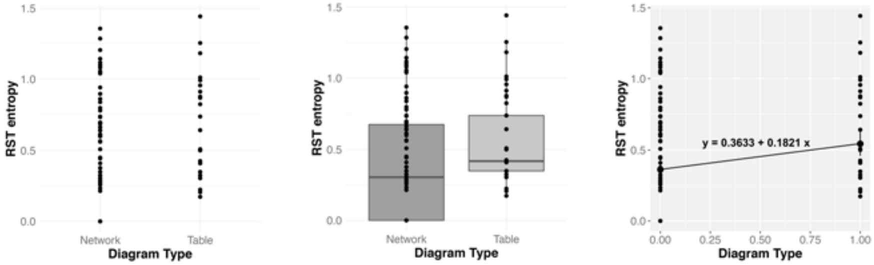


Fig. 2: From data to correlation to difference: data extracted from the AI2D-RST corpus for 179 diagrams; left: raw data, middle: raw data with descriptive statistics visualizations; right: raw data with regression line.

spread of the values: the taller the boxes, the more spread out the values are. Since the two boxes are quite different in height and only overlap vertically to a limited extent, this could well be indicative of a significant and reliable difference between the complexity of the two diagram types. But, again, ideally we would need to perform some statistical test that would either lend support to this impression or show that we do not yet have evidence of a reliable difference after all.

As described for multimodality research more broadly in Bateman et al. (2017), there are a considerable number of statistical tests available and which tests one should select to show a difference depend on a variety of properties of the data. Our goal here will be more specific, however. Following the method set out by Field et al. (2012), we can relate the example research question directly to the formation of a generalized linear model simply by calculating a regression line for the data exactly as was shown in Section 4.1. Just as before, we have in the present case a collection of pairs of values that define points on a graph. The vertical axis values are given by the rhetorical complexity values, while the horizontal axis values are given by the visual genre categories. However, different to the previous case, there are only two horizontal axis values that ‘make sense’, or are defined, one for ‘Networks’ and one for ‘Tables’. Nevertheless, once we have a collection of points on a graph, we can calculate regression lines exactly as before. In fact, we could in principle generalize this to any number of qualitative categories; all that is required is that the categories are mapped to distinct positions in the graph.

The result of performing a regression calculation for the two categories case is shown on the right of Figure 2, where the line connecting the two vertical groups of points represents the calculated regression line. The equation describing the line is also shown as before: for this calculation it makes no difference that the points were grouped together vertically rather than spread over the entire graph as was the case in Figure 1: the line shown still minimizes the collective distance between the line and the points. Finally, the line, or rather its equation, also lets us calculate a prediction of rhetorical complexity given a value for the visual genre — even though there are now only two values along the horizontal axis defined as meaningful: the value zero, set here to correspond to ‘networks’, and the value 1, used for ‘tables’.

Generating this model in R is done in precisely the same way as before. With the labels of the data points adapted accordingly we write:

```
glm( formula = RSTEntropy ~ DiagramType, data = diagram_data )
```

This instructs R to construct a model whose task it is to describe the relationship between the `RSTEntropy` (i.e., the measure of rhetorical complexity) and `DiagramType` (the visual genres) as given in the diagram data. When we ask R to summarize the constructed model so that we can evaluate the results for our research question, we obtain the following (abbreviated) output:

```
Coefficients:
                  Estimate    Std. Error    t value    Pr(>|t|)
(Intercept)        0.36330       0.03112      11.67     < 2e-16 ***
DiagramType.Table  0.18210       0.06010       3.03     0.00281 **
```

Just as above, this table reports the coefficients of the linear model, that is, the numbers that define the regression line shown, and the respective values indicating how confident we should be that the result is reliable or not. As before, the `Estimate` column tells us how to draw the regression line on the graph: the `Intercept` is where the line crosses the zero value on the vertical axis, which in the current example corresponds to the visual genre of ‘networks’, and the `DiagramType.Table` tells us the slope of that line as indicated in Figure 2 (i.e., 0.1821). Now, since the visual genre ‘network’ is associated with the value zero, this genre is used as the reference value. The precise meaning of the slope described for `DiagramType.Table` is then that when moving from the reference genre (‘Network’ or zero) to the identified genre (‘Table’ or 1), the slope is as given. Since the distance between the visual genres is 1 unit, this slope means that the value of rhetorical complexity for the ‘table’ visual genre is then predicted to be 0.3633 (the intercept) plus 0.1821 (i.e., the slope times the number of units moved, i.e., 1). This value is 0.5454, which turns out to be the average value of rhetorical complexity for the visual genre ‘tables’ in the data — that is, the model predicts that the rhetorical complexity for tables will be the average of the values it has seen in the data, which is quite a reasonable assumption.

The question then remains as to whether this is a reliable, i.e., statistically significant, difference between the two visual genres. The summary table above shows us that this is indeed the case. The column `Pr(>|t|)` gives the probability that the observed change in the value for rhetorical complexity across the two genres would have occurred under the assumption *that the two groups exhibit no difference in rhetorical complexity*. In other words, this is the probability that the *range* of values expected for the slope includes zero (i.e., there is no change when going from one group to the other). In traditional statistics, this model corresponds to what is called the 'null hypothesis', or the assumption that there is no difference in the data to be observed. One then calculates how well this most simple model would predict the data observed: that is, assuming that the model is an accurate description of the data, how well does it do? The value calculated for this probability for the slope is shown in the last row and column of the summary table to be 0.00281. This probability is quite low (and less than the often assumed cut-off point for statistical significance of 0.05), and so we can confidently reject this possibility: it is *not* likely that the estimate of the slope when considering some other sample of data instead of the present one will include the value zero, i.e., no change between diagram types. Any model that predicts no difference would not then be a good choice. This gives us the answer to our initial research question: on the basis of the data we can state that there does appear to be a reliable difference between the rhetorical complexity of the visual genres considered.

Linear models gain their wide applicability for predictive modeling by virtue of their foundation in inherently generic statistical methods — methods which are precisely geared towards detecting and interrogating patterns. Moreover, the close relationship indicated between modeling and prediction allows seeing statistical tests from two complementary angles: indeed, any statistical test can in fact be re-construed as a model. We can suggest this further by considering how we might have answered our research question concerning visual genres using a more traditional statistical test. Since we are dealing here with two qualitatively defined groups of data (the visual genres) and a continuous value (the complexity), the standard statistical test that would be used for ascertaining whether the two groups are significantly different or not is the t-test (cf. e.g., Bateman et al. 2017: 178-184). When we run the t-test on the current data with any standard statistics tool, we obtain a value for the probability of the data as observed occurring given an assumption that there is no difference between groups (i.e., the null hypothesis). If this probability is sufficiently low (again, usually less than 0.05), then we can reject the assumption that there is no difference between the groups and come to the conclusion that, in all likelihood, the two visual genres indeed differ in terms of their rhetorical complexity. In fact, for the current data, running the t-test reports an associated probability of this configuration occurring if the two groups were the same of 0.00281. This is precisely the same value as was produced using the generalized linear model above, showing the close relationship between the two approaches.

This is an important interim result because it begins to make clear how constructing generalized linear models for data in fact renders many situations amenable to a completely uniform treatment. Typically, when we change the form of our data, moving from single variables to several variables, or from single groups to several groups, we have to look for different statistical tests as well. This is because traditional statistical tests are tightly defined to fit to very specific forms of data. In contrast, the generalized linear model can be used for quite different kinds and distributions of data. All that varies is the complexity of the predicting formula and the kinds of statistical distributions allowed when fitting a model's coefficients, which can all be adapted quite simply by changing the information we pass on to R's `glm` function. To show this, in our final example we make our generalized linear models more complex by introducing additional *predictors* to the model. Whereas before we examined the potential relationship between two variables (i.e., the RST complexity and the type of diagram), now we explore to what extent the rhetorical complexity might *also* be influenced by the quantity of text and graphical materials present in each diagram.

This rather complex situation is straightforwardly encoded by adding the number of `Text` and `Graphic` elements in each diagram into our generalized linear model formula as predictors of rhetorical complexity as follows:

```
glm( formula = RSTEntropy ~ DiagramType + Text + Graphic, data =
                           diagram_data )
```

This instructs R to produce a model that considers to what extent the independent variables `DiagramType`, `Text` and `Graphic` may contribute separately to the prediction of `RSTEntropy`. A summary of the model that R constructs yields the following coefficients:

```
Coefficients:
                  Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)       0.124155     0.058262     2.131   0.034486 *
DiagramType.Table 0.150264     0.060051     2.502   0.013257 *
Text              0.019794     0.005467     3.621   0.000384 ***
Graphic           0.008364     0.004920     1.700   0.090941 .
```

This table is to be read in precisely the same way as the table above; there are simply more coefficients listed because we are examining the contributions of more potential predictors. This shows concretely how we can progressively increase the complexity of our models while using the same overall modeling strategy.

To interpret this table, we can begin as before by considering the contribution of the type of diagram, which is given by the row of the table labelled `DiagramType.Table`. Interestingly, the estimate for this predictor, i.e., the estimate for the slope when moving from the visual genre of ‘Network’ to that of ‘Table’, is now given as 0.15 — recall that above we obtained the value 0.18. This difference in result is caused by the fact that our model now includes several further potential predictors and these appear to be making their own contributions to predicting values for rhetorical complexity. In other words, the work of prediction is now being spread over the new predictors as well. Note that these new predictors have not taken over all the work since the table shows that the diagram type still makes a statistically significant contribution to rhetorical complexity. This is given by the final value in the row, which is still sufficiently low (0.0133) as to reject the hypothesis that the calculated slope might by chance include zero, i.e., that there is no reliable contribution for diagram type.

The following row then shows us that there is an extremely significant effect contributed by the `Text` predictor. Rhetorical complexity is therefore also associated with the quantity of text in a highly reliable fashion. This might well have been expected, as more text gives more opportunities for rhetorical complexity, but the results here allow us to turn this expectation into a reliable prediction. Furthermore, just how the quantity of text and the prediction of rhetorical complexity are related is given by the number in the Estimate column. In contrast, the number of `Graphic` elements does not reliably predict an increase in rhetorical complexity to a statistically significant degree (the full stop next to the probability value indicates a value between 0.05 and 0.1, which is not generally a level of significance that would be considered reliable).

We have now shown how one can make models more complex, but this naturally raises the question of just how complex we *need* to make our models. The general rule is, as might be expected, that models should be as complex as necessary but not more so. And here the form of modeling we have presented also provides direct help. It is possible to calculate for each increase in complexity of a model whether the performance of the more complex model is statistically significantly better, i.e., does a better job at predicting values, than the less complex model. If a more complex model does not deliver better results, then there is no need to adopt it. Several techniques exist for comparing series of models that have been made progressively more complex. When producing the models with the R `glm` function above, for example, an additional value is reported that offers a measure of the ‘goodness’ of the model produced: this is the Akaike Information Criterion (AIC). The definition of this value need not concern us here — important is only that ‘lower values are better’ when comparing models: i.e., the smaller the AIC, the better the fit between model and data.

The AIC value that R reports for our first example above involving just rhetorical complexity and diagram type is 142.41. This tells us nothing in itself, but the AIC reported for the more complex model that we have just interpreted, with the quantities of text and of graphics included as well, is 123.67. This is less than the previous value and so we know that the second, more complex model fits the data better than the first model and so would be preferred. Crucially, however, adding predictors does not always lead to an improvement. Thus, for example, if we were to add a further predictor for our data, such as the number of arrows found in each diagram analyzed, the new model generated tells us that diagram type and text both help predict the rhetorical complexity as before, but graphics and the added predictor of 'arrows' do not; we omit the generated table of coefficients here for reasons of space as it looks precisely like the table above simply with another row. Now, the AIC reported for this last, even more complex model is 124.74. This is *greater* than the AIC reported for the previous model and so we know that making the model more complex in this way is not warranted. At that point, therefore, we could stop our modeling process knowing that, for the time being and until more data or predictors become available, we have done the best we can.

In general, following such a method one can pursue model building quite experimentally, progressively trying out (theoretically motivated) predictors and only keeping them in a growing model when they actually make a difference to the model's performance. Expressing models in terms of generalized linear models of the kind used here makes this process quite straightforward as we have seen. For each model produced, one can also investigate interpretations of the model by turning it 'around' and using it for prediction: that is, once we have a model, that model can be used to predict outcomes given any combination of values for the predictors, including combinations that did not occur in the dataset. This offers yet another way of exploring the consequences of models and considering whether they are also theoretically well-founded. In all cases, one is using properties of the models to find out more about their object, i.e., the data being modeled, and so are functioning as iconic signs just as argued above.

5 Conclusion

In this chapter, we have discussed how the empirical foundations of multimodality research can be strengthened by paying increased attention to the semiotic nature of models and the growing range of techniques for formulating models and evaluating them against data. All multimodal researchers who have taken the step to work with larger bodies of material face the challenge of finding patterns in the observations they make of that material. This raises many substantial issues in its own right and is one of the reasons why orienting towards a broader array of potential methods is now crucial. Although it has long been suggested that patterns in multimodal data may be revealed by visualization (cf. Manovich 2011; O'Halloran et al. 2012; Caple et al. 2018), a more thorough semiotic grounding in iconicity reveals that modeling is essential here too. Visualizations demand interpretations and such interpretations are themselves models, albeit often implicit ones. There is, therefore, much that needs to be done to improve the situation, particularly for relating visualization and other more 'bottom-up' forms of data processing to modeling — topics which space has precluded us from addressing here. In general, however, it is unlikely that the complexity of interrelationships at work in multimodal communication will be revealed without employing more sophisticated techniques for finding patterns in data such as those this chapter has begun to introduce and we hope that our discussion has shown how the role of models in empirical research can be substantially extended. Model-building should then, we believe, be a standard component of evaluating empirical results from all kinds of multimodality studies.

Bibliography

Ambrosio, Chiara. 2014. Iconic Representations and Representative Practices. *International Studies in the Philosophy of Science* 28(3). 255–275. https://doi.org/10.1080/02698595.2014.959831.

Bateman, John A. & T. Hiippala. 2020. Statistics for Multimodality: Why, When, How – An Invitation. SocArXiv https://doi.org/10.31235/osf.io/7j3np.

Bateman, John A., J. Wildfeuer & T. Hiippala. 2017. *Multimodality – Foundations, Research and Analysis. A Problem-Oriented Introduction*. Berlin: De Gruyter Mouton.

Box, George E.P. 1979. Robustness in the Strategy of Scientific Model Building. In R. Launer & G. Wilkinson (eds.), *Robustness in Statistics*, 201–236. Academic Press. https://doi.org/10.1016/B978-0-12-438150-6.50018-2.

Caple, Helen, M. Bednarek & L. Anthony. 2018. Using Kaleidographic to Visualize Multimodal Relations Within and Across Texts. *Visual Communication* 17(4). 461–474.

Cohen, Jacob. 1968. Multiple Regression as a General Data-Analytic System. *Psychological Bulletin* 70(6, Pt.1). 426–443. https://doi.org/10.1037/h0026714.

Field, Andy, J. Miles & Z. Field. 2012. *Discovering Statistics Using R*. London: Sage.

Fricke, Ellen. 2012. *Grammatik multimodal: Wie Wörter und Gesten zusammenwirken*. Berlin and New York: De Gruyter Mouton.

Hiippala, Tuomo, M. Alikhani, J. Haverinen, T. Kalliokoski, E. Logacheva, S. Orekhova, A. Tuomainen, M. Stone & J. A. Bateman. 2020. AI2D-RST: A Multimodal Corpus of 1000 Primary School Science Diagrams. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-020-09517-1.

Hookway, Christopher. 2000. *Truth, Rationality, and Pragmatism*. Oxford: Oxford University Press.

Jewitt, Carey (ed.). 2014. *The Routledge Handbook of Multimodal Analysis*. London: Routledge 2nd edn.

Kendon, Adam. 2004. *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Kralemann, Björn & C. Lattmann. 2013. Models as Icons: Modeling Models in the Semiotic Framework of Peirce's Theory of Signs. *Synthese* 190(16). 3397–3420.

Kress, Gunther & T. van Leeuwen. 2001. *Multimodal Discourse: The Modes and Media of Contemporary Communication*. London: Arnold.

Kress, Gunther & T. van Leeuwen. 2006 [1996]. *Reading Images: The Grammar of Visual Design*. London and New York: Routledge.

Manovich, Lev. 2011. What is Visualization? *Visual Studies* 26. 36–49.

Martinec, Radan & A. Salway. 2005. A System for Image-Text Relations in New (and Old) Media. *Visual Communication* 4(3). 337–371.

Norris, Sigrid (ed.). 2016. *Multimodality; Critical Concepts in Linguistics*. Amsterdam: John Benjamins.

Nöth, Winfried. 1995. *Handbook of Semiotics*. Bloomington: Indiana University Press.

O'Halloran, Kay L., A. Podlasov, A. Chua & M. K. E. 2012. Interactive Software for Multimodal Analysis. *Visual Communication* 11(3). 363–381. https://doi.org/10.1177/1470357212446414.

Peirce, Charles Sanders. 1998 [1893–1913]. *The Essential Peirce – Volume 2. Selected Philosophical Writings (1893–1913)*. Bloomington: Indiana University Press.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/ (last accessed: 1 September 2021).

Schreier, Margrit. 2012. *Qualitative Content Analysis in Practice*. London: Sage.

Serafini, Frank & S. Reid. 2019. Multimodal Content Analysis: Expanding Analytical Approaches to Content Analysis. *Visual Communication* aop. https://doi.org/10.1177/1470357219864133.

van Leeuwen, Theo. 2005a. *Introducing Social Semiotics*. London: Routledge.

van Leeuwen, Theo. 2005b. Three Models of Interdisciplinarity. In R. Wodak & P. Chilton (eds.), *A New Agenda in (Critical) Discourse Analysis: Theory, Methodology and Interdisciplinarity*, 3–18. Amsterdam: John Benjamins.

Winter, Bodo. 2020. *Statistics for Linguists. An Introduction Using R*. New York and London: Routledge.

Barbara Tversky and Angela Kessell

Thinking in Action

Reprint

Abstract: When thought overwhelms the mind, the mind uses the body and the world. Several studies reveal ways that people alone or together use gesture and marks on paper to structure and augment their thought for comprehension, inference, and discovery. The studies show that the mapping of thought to gesture or the page is more direct than the arbitrary mapping to language and suggest that these forms of visual/spatial/action representation are used to "translate" language into mental representations. It is argued that actions in space create patterns in the world that reflect abstractions, that the actions are incorporated into gestures and the patterns into diagrams, a network that integrates gesture, action, the designed world, and abstraction dubbed *spraction*.[1]

Keywords: gesture, diagram, visual communication, design, thinking, action, language

1 Thinking in Action

How does thinking happen? How does the mind talk to itself? One view is that the language of thought is exactly that, language (narrow), and that language appeared suddenly in human history to serve that function (see Chomsky in Gondry 2013; see also Fitch et al. 2005). Yet, watching people think has become a common sight in the era of cell phones and that spectacle suggests other ways that thinking happens. What is fascinating, and often comic, are the gestures people make as they hold their small rectangular blocks speaking into the air. Those gestures cannot be intended for their remote addressees; rather they seem to serve speakers. Because speakers lose fluency when prevented from gesturing, some have proposed that gesturing for self helps speakers find words (e.g., Krauss et al. 2000; Rauscher et al. 1996). However, we are finding, and will relate below, that people gesture

[1] This chapter first appeared as Tversky, B. and Kessell, A. 2014. "Thinking in action" In: *Pragmatics & Cognition*, 22.2, 206-223. It was published as part of a special issue on "Diagrammatic Reasoning", edited by Fusaroli, R. and Tylén, K., published by John Benjamins, Amsterdam/Philadelphia. We are grateful to both the authors and the publisher for granting us permission to reprint the contribution in our volume, and to Barbara Tversky in particular for delivering an insightful keynote lecture on the topic at *BreMM19*.

https://doi.org/10.1515/9783110725001-004

in the absence of speaking, and that these gestures express thought, abstract thought, structure thought, and augment thought. When given props, salt and pepper shakers, silverware, pen and paper, people use the props for similar ends. They use these cognitive tools, their bodies and props, to think and communicate, not only for themselves and but also for others. This, along with findings of many others (e.g., Donald 1991; Goldin-Meadow 2003; Hutchins 1995; Kirsh 1995; Norman 1993; Suchman 1987), is good evidence that people use their bodies and the world to think.

Gestures and arrangements of props, sketches, and diagrams are all forms of visual thought, of visual communication. They work, but they work differently from language, more directly than language. Compare a map to a verbal description of an environment. A map "maps". It takes objects and spatial relations in a real world and puts them on a virtual world, typically a page, preserving certain spatial (and sometimes visual) properties like distance, size, and direction. Typically maps omit certain information and exaggerate other information as well, depending on their intended use (e.g., Tversky 2011). Language can represent an environment but it does not preserve spatial properties directly. Gestures can create maps that preserve spatial properties (e.g., Emmorey et al. 2000). Diagrams can map abstract objects and relations to marks and spatial relations on a page, as in corporate charts or decision trees. Even a shopping list, which seems to primarily serve to off-load memory, is more effective if the items are listed in the order of navigating the supermarket, that is, if the list is structured in a cognitively meaningful way. Externalizing thought to a page has clear benefits for memory, thinking, inference, and discovery (e.g., Tversky 2011). Here, we will explore some of the ways that gestures can do the same, and to draw parallels between these forms of visual thought and communication.

We begin by describing research in which participants alone in a room spontaneously gesture or sketch to support their thinking in a range of tasks, showing that doing so enhances their thinking. We continue by describing similar phenomena in social situations, illustrating that some of the ways the mind talks to itself resemble the ways that minds talk to each other, and suggesting, echoing Vygotsky (1962), that communicating, with words, hands, or props, to one's self can be viewed as internalized social interactions. We argue that this form of thinking is both disembodied and embodied: internalized perception, action, and interaction inside the mind, reexternalized outside the mind. We show that gesture and sketch can represent thought more directly than symbolic words, yet, like words, can abstract. Then we show that gesture, sketch, and abstraction interact in a network, termed *spraction*, that integrates actions in space, consequential patterns in space, and abstraction.

We begin by describing some experiments, primarily our own, and the phenomena they elucidate. The first experiment is described in greater detail here because the details do not appear elsewhere.

2 Gesturing and Sketching for Problem Solving

Problem solving is widely regarded as having two stages, understanding the problem, and finding a solution (e.g., Holyoak 1994; Newall et al. 1958). Understanding a problem entails abstracting the essential elements and their relations, thereby forming a mental representation or mental diagram of the problem (see Stjernfelt 2011 on Peirce; Tversky 2011). Solving a problem typically entails interacting with the elements, that is, mental actions. Creating real diagrams can facilitate problem solving, simple ones like multiplication and difficult ones like designing buildings or understanding the interactions of tiny particles or heavenly bodies. The real diagrams serve to reduce memory load by putting it into the world. However, good diagrams also serve to structure a problem, to extract the essential information, to omit the irrelevant, and to structure the relevant (e.g., Tversky 2011; Zhang & Norman 1994). In multiplication, for example, this means not just putting down the numbers to be multiplied which would off-load memory, but also lining up the columns by 1's, 10's, 100's, which structures the problem appropriately. Compare Roman and Arabic numerals; both off-load memory, but performing arithmetic operations is far easier with Arabic numerals Zhang & Norman (1994).

In the absence of paper, would problem solvers create virtual diagrams with their hands to explain problem solutions, and perhaps even to solve them? Would they use their hands as they use paper? In a pair of experiments, Kessell & Tversky (2006) asked 44 Stanford undergraduates to solve and then explain six insight problems. Half the participants had paper and pen and half did not. Participants were alone in a room; a camera recorded the sessions. They were given four minutes to solve each problem. After they solved the problems, they turned to face the camera to explain their solutions, correct or incorrect, so that someone else viewing the video could understand the problem and its solution. Some problems, like the well-known Radiation (Duncker 1945) and Two Strings (Maier 1931) problems required keeping track of only 2–3 separate elements, a relatively small working memory load, but two problems, Glasses and Ladder, required keeping track of 6 or more elements, a load likely to tax working memory. Our discussion will concentrate on those problems. For *Glasses*, participants were told there were three full and three empty glasses in a row; they were asked to change the array to alternating empty and full glasses by moving a single glass. Most participants correctly solved the problem, with the insight of pouring the liquid from the middle glass of the row of full glasses into the middle glass of the row of empty glasses. *Ladder* was a red herring: a ladder with 10 rungs separated by specific distances hung from a boat with the water reaching the bottom rung, the tide was rising at a specific rate, and the question was about the expected water level relative to the ladder rungs after a specified time. Because boats float, the water level with respect to the ladder did not change as the tide came in. However, most participants neglected the fact that the boat floats and computed the rising level as if the boat were attached to the bottom of the sea rather than floating. Representational gestures, iconic and deictic, were coded, most by two coders with high agreement (Cohen's $K = 0.90$). Many gestures are both; e.g., for Glasses, many people pointed to hypothetical glasses, a deictic gesture. However, the gestures were horizontally arrayed, an iconic correspondence.

Surprisingly, a majority (62% for Glasses, 64% for Ladder) of participants without paper gestured while they solved the problems, alone in the room, without speaking. Very few gestured for the problems with few elements, a difference that was significant ($\chi_1^2 s > 8.76$; $ps < .01$). Similarly, those with pen and pencil diagrammed Glasses and Ladder, but far fewer diagrammed the other problems ($\chi_1^2 s > 5.35$; $ps < .05$). Both the gestures and the diagrams abstracted the essentials of the problems and represented their structure, the two sets of three glasses arrayed horizontally or the 10 rungs of the vertical ladder. Gestures also frequently enacted solutions, pouring for Glasses and counting rungs for Ladders. Correspondingly, the diagrams included arrows for Glasses and numbers for Ladders. Thus, both gestures and diagrams were used not only to off-load memory but also to abstract and structure the problem, eliminating unnecessary features of the described situation, such as the boat, but adding key features not in the problem description, such as arrows. As noted, a large majority of participants solved Glasses and failed Ladder. For Ladder, a majority correctly computed the answer under the incorrect assumption that the boat was anchored to the ocean floor. Interestingly, participants who gestured were more likely to solve Glasses than those who didn't gesture. For Ladder, those who gestured were more likely to correctly compute the incorrect solution to Ladder than those who didn't gesture. These results did not quite reach significance because so few failed to solve Glasses or to solve Ladder with the correct insight.

Gesturing while explaining was quite different. All the problems elicited gestures from nearly everyone. The forms of the gestures for explanation were quite similar to the forms of gestures for self, for example, points and pours for Glasses and flat hands and counting for Ladder. However, there were more gestures in explanations to others, and the sequence of gestures formed complete narratives, laying out the problems and enacting their solutions (as in Emmorey et al. 2000). That is, in explanation, the gestures modeled the problem structure and solutions.

Importantly, the gestures expressed nuanced differences in meanings across participants, despite similarities in words. For Glasses, some pointed at places, suggesting locations; others made grasping gestures, suggesting movability. One wonders if those who make grasping gestures would solve the problem more quickly. For Ladder, some made continuous gestures to indicate the rising tide and others made discrete gestures at each imagined rung. The hands shifted meaning in some cases, seamlessly. For example, in the two strings problem, first both hands represented the two ropes hanging vertically, then they represented hands, picking up the hammer and tying it to the end of one rope, then a single hand represented the swinging action of the rope, and finally, the hands represented hands again, tying the two ropes together.

Like the gestures, the sketches seemed to serve different roles in thinking and in communicating. Here the sketches primarily served thinking. Only 32% of those who had diagrammed Glasses and 21% of those who had diagrammed Ladder used their sketches in their explanations, lifting them and turning them toward the camera. Nearly all participants gestured in explaining nearly all the problems, even when they had made diagrams, with about half the gestures on the diagrams and half in the air.

Both gesturing and sketching for self went beyond externalizing memory. They didn't simply put the elements on paper to extend working memory. The gesture and diagram arrays represented the structures of the problems and their solutions, that is, the organization of the elements and the relations described by the text, and the spatial/enactive forms of the solutions. The gestures and sketches embodied participants' understandings of the problems, extracted from the text, but not the text per se. Both gestures and diagrams abstracted the problem descriptions; that is, they included only the details important for representing and solving the problems, leaving out the information irrelevant to the solution; e.g., for Ladder, the ladder but not the boat. They also added information that was key to solutions but not in the problem descriptions. That is, participants' gestures and sketches selected information and reconfigured it to represent their understandings of the problems. Consequently, those representations should affect the search for solutions by directing the search. Altogether, it appears that creating spatial-motor representations through gesturing or sketching helped to establish mental representations and to facilitate finding solutions (correct or not).

Despite these similarities in the use and structure of gestures and sketches in problem solving for self and in explanation for others, there were important differences as well. Communicating with one's self for problem solving, or for anything, can be and is more telegraphic, as the self knows more, notably the surrounding context, details, and interconnections. We have more common ground with ourselves than with others. Communicating with others has to establish that context and to present a complete and coherent explanation. Thus, communicators used more gestures than problem solvers. There were also inevitably differences between gesture and diagram, but discussion of that will come later.

3 Gesturing for Understanding and Remembering

The problem solving study suggested that gestures for self help solving problems, but more evidence is needed to establish that gesturing can help one's own thinking. Describing space to others elicits copious gesturing, even in the blind (e.g., Iverson 1999). If describing space to others elicits gestures, perhaps learning spatial environments would elicit gestures for self. To this end, we adapted spatial descriptions from previous work Taylor & Tversky (1992). The descriptions were from either route or survey perspective. Route perspectives take imagined travelers on tours of the environment, describing landmarks relative to the travelers' embedded point of view, using *right*, *left*, *front*, and *back*. Survey perspectives take an overview perspective and describe landmarks relative to each other, using *north*, *south*, *east*, and *west*. In this experiment, 48 participants alone in a room studied four spatial descriptions to prepare to answer later true-false questions (Jamalian et al. 2013). Half the descriptions were of small (4 landmarks) environments and half of large environments (8 landmarks), half from each perspective. Similarly, half the true-false questions were from each perspective, irrespective of the read perspective. Some questions were taken verbatim from one of the texts and others required inferences from the information in the text. As in previous research, memory performance was good but not perfect. Read perspective made no difference in performance, but accuracy was higher for verbatim rather than inference statements.

Surprisingly, about 70% of participants spontaneously gestured at study for at least one description, and most of those also gestured when they answered questions. Gesturing helped. Those who gestured at learning or at test performed considerably better than those who did not gesture. Furthermore, those who gestured on some but not all of the descriptions or questions performed better on the descriptions or questions for which they gestured. Gesturing was self-determined, so an experiment allowing or not allowing gesture is in progress.

As before, the gestures appeared to structure memory over and above offloading memory. The rate of gesturing did not depend on size of environment; it was the same for small environments, likely to be within working memory capacity, and for large environments, which likely exceeded working memory capacity Although the descriptions were full of rich visual information, participants gestures overwhelmingly represented only the spatial structure of the environments. They represented paths with line-like gestures of diverse forms and the locations of landmarks with point-like gestures, analogous to spontaneously produced sketch maps, which consist primarily of dot or blob like marks for landmarks or intersections and lines for paths (Tversky & Lee 1999). The same gestures were often repeated, as if to emphasize, to stamp into memory, analogous to overdrawing.

Other researchers have found similar results, that participants' gestures appear to represent thought and to facilitate it, for mental rotation (Chu & Kita 2008; Schwartz & Black 1996; Wexler et al. 1998; Wohlschläger & Wohlschläger 1998), for counting (Carlson et al. 2007), for math and more (Goldin-Meadow 2003; Goldin-Meadow et al. 2009; Hoeststetter & Alibali 2008).

In our experiments, participants rarely looked at their hands, only brief glances now and then. This strongly suggests that the gestures created spatial/action representations rather than visual ones, and that the facilitation was spatial/action rather than visual. Think of locating the X key on a typewriter, or reaching for the light switch in the dark. Think of skilled musicians, playing exquisitely, looking at the score. Think of navigation and action in the blind. What is remarkable in the studies described is that participants used actions of their hands to represent information that is not experienced by their hands, that is, the knowledge is represented in the actions of the hands.

4 Sketching for Discovery in Design and Art

We have established that people spontaneously gesture and sketch to comprehend, remember, and think, that their gestures and sketches represent the structure of thought, not the words they have heard or read, and appear to aid thinking. The cases we just considered are cases that require focusing thought to a single clear uncluttered interpretation, *the* solution to a problem, *the* structure of an environment. There are other cases as diverse as design, innovation, art, and data analysis where thought begins and often remains inchoate, emergent, unclear, ambiguous, and where these qualities are desirable, as they allow for interpretation and reinterpretation, for seeing alternatives, for avoiding fixation. Schon has famously described designers as having conversations with their sketches (Schon 1983).

Design. We studied the conversation designers have with their sketches in expert and novice architects (Suwa & Tversky 1996, 2003; Suwa et al. 2001; Tversky & Suwa 2009). In the first project, experienced and newly-minted architects were asked to design a museum, with certain constraints. They were videotaped as they designed, and afterwards, viewed the videos with the experimenter, explaining why they drew each mark, what they were thinking. Notably, the architects make discoveries in their own sketches. That is, they drew a set of marks for one reason, but on inspection, saw new and unintended aspects in their own sketches. The experienced architects made more discoveries in their own sketches than the novice architects. Especially for the experienced architects, the new discoveries led to positive cycles, new sketches of new ideas, and more new discoveries. What's more, the experienced architects made more conceptual inferences than novices, who made primarily perceptual inferences. A perceptual inference could be more or less read from the sketch, for example, seeing a pattern in the array of buildings, and then using that pattern as a motif. A conceptual inference required information not in the sketch, for example, seeing how traffic would flow or how lighting would change over the seasons. Ambiguity allowed the reconfigurations that supported reinterpretations, and knowledge and experience allowed for richer ones.

We developed a simpler task that allowed seeing reinterpretation in action (Suwa & Tversky 2003). In several experiments, designers, architects, and ordinary people were shown ambiguous sketches repeatedly and asked to come up with a new interpretation each time they saw the sketch. Architects and designers came up with more interpretations than ordinary people. Those producing more interpretations reported reconfiguring the elements of the sketches to do so. Again, ambiguity allowed reconfiguring. An experienced architect facile at new interpretations in the previous study reported the same. We found that people facile at detecting small geometric figures in large complex ones (Embedded Figures) were better at the task, as were people facile finding remote associates, and that these abilities, one perceptual, one conceptual, were independent. We proposed that the process of finding new interpretations in ambiguous sketches is one of *Constructive Perception*, the active use of perception in the service of innovation. Thus, the conversations people have with their sketches can be nurtured and trained. Interestingly, when architects are asked to design blindfolded, they make copious gestures (Bilda & Gero 2006).

Art. A similar conversation occurs between artists and their sketches, even when the sketches are an end in themselves rather than a representation of something to be constructed. Kantrowitz (2014) adopted Suwa and Tversky's methodology in her study of 8 experienced artists for whom sketching is a major part of their practice. Although most of them begin with a vague plan, after starting, they look at their sketches to see how to continue. As one said, "I want the work to tell me what it wants (p. 136)." Others observed that drawing allows them to enter and explore a new world, even "intentionally getting lost (p. 106)." They point to the emergence of configurations from features. One said, "I see a form emerging from the sort of more random marks" and later, "It's almost as if the form has appeared on its own (p. 148)." Significantly, here and in other research on drawing, artists find it distracting to talk while they draw. One observed, "the hand tells the eye where to go and the eye tells the hand when to stop (p. 114)." That is, the dialogue is visual/spatial/action/; it is between the eye and the hand and the sketch. It surely involves the mind, but it is not mediated by words; they seem to interfere.

5 Gestures and Sketches for Others

We have described many ways that people use forms of visual/spatial/action communication to talk with themselves. There are by now a multitude of studies showing that people use gesture, body language, sketches, artifacts, arrangements of the world in their comunications with themselves and with others, several described in other papers in this issue (for more, see Clark 1996; Goldin-Meadow 2003; Hutchins 1995; Kendon 2004; Kirsh 1995; McNeill 1992, 2005; Norman 1993). Here, we briefly describe another of our own, exploring how gesture, diagram, and language are integrated in collaborations (Heiser et al. 2004). Thirty pairs of Stanford students were given a map of campus after a hypothetical earthquake, showing roads blocked and number of wounded at various points. Their task was to find the best route to collect as many wounded as possible as quickly as possible, and draw the route on another sheet of paper. There was more than one good solution, and most dyads finished the task in less than 25 minutes. Half the pairs, the *remote* dyads, were separated by a shower curtain; each had a copy of the original map, but they had to negotiate the rescue route only using speech. After negotiating and agreeing on a route, each of the pair drew a map of the rescue route. The other pairs were *co-present*. The co-present dyads shared the original map, and drew a single rescue map.

The conversation of the co-present group was conducted primarily as a gestural narration on the given map. Participants looked at the gestures on the map, not at each other. The set of gestures was small, primarily points for locations, for example, of turns or of places where wounded were grouped, lines for paths, and whole hand swipes for regions. We have already seen gesture points for places and lines for links in the experiment in memory for described environments. Here, the gestures represented concepts that were conceived as 0, 1, or 2 dimensional, points, lines, and sweeps for regions (cf. Talmy 1983: 200). Interestingly, many of the remote pairs gestured on their maps when listening to their partner's suggested route, following the route with their fingers on the map, to understand that route. In the co-present dyads, participants' gesturing on the maps took turns, in tandem with the speech, and the current speaker controlled both. Sometimes you could see the listener's hand at the edge of map, waiting eagerly to get in. As the collaboration progressed, their gestures were entrained and abbreviated. The co-present dyads were more collaborative than the remote dyads, they had more equal contributions, took turns more often, and worked on the same task at the same time. They reported being happier with the experience on a number of measures. Significantly, despite reaching agreement, the routes produced by the remote pairs differed 30% of the time! Participants thought they had agreed when they hadn't. That is, the spoken language was not sufficient to clarify and disambiguate the routes. Gestures on diagrams, with each other and with words, more easily clarify and disambiguate. It is clear that the gestures contributed to the joint problem solving, that the gestures extracted the essential information from the maps and structured it into proposed routes. It is notable that the gestures on the maps are similar to the gestures that were used to create spatial/action representations in the absence of a map. The same gesture vocabulary was used, points for places, lines for paths. Thus, there are striking parallels in communication with self and communication with others.

6 How Do Gesture and Sketches Help Thought?

It is clear that people spontaneously gesture and sketch to talk to themselves, to think actively and productively, to comprehend, structure, remember, draw inferences, and make discoveries. They use the same media in talking with others. In fact, it seems that a critical role for gestures and sketches is to translate words and sentences into thought. The implication is that the words and sentences can need translation, that they are not always sufficient for thinking. As noted, the structure of language and the structure of visual forms of communication differ. Words come one after another in a more or less linear order; that order is governed for the most part by syntax and pragmatics. The arrangement of words only partially corresponds to the arrangements of the worlds the words describe. Notably, events are often, though by no means always, described in the order they occur. Order is a weak form of structure, not sufficient to correspond to the complex 2- or 3-dimensional structures that marks on a page or gestures in the air can represent (but see Netz 1999)). Comprehension requires creating mental models, not just of content that is inherently spatial, but also content that is abstract, including discourse (e.g., Kintsch 1996), science and engineering systems (e.g., Gentner & Stevens 1983), and logical arguments (Johnson-Laird 1983). Externalizing, or reexternalizing, mental models in corresponding gestures and diagrams appears to help the formation of mental representations. Thus, representing many ideas in a visual/action/spatial medium is more direct and natural than representing them in a purely symbolic medium such as language. How does this medium work?

Creating a mental model or a mental representation entails selecting and organizing key information from the external world and mapping that to an internal representation. That is, some information is included and some discarded. It typically entails adding Information from previous knowledge. Mental representations or models are key to comprehension, which in turn is key to reasoning, to going beyond the information given (Bruner 1957). The core of such representations is the elements, abstract or concrete, and the relations among them, abstract or concrete; the relations form a skeleton or framework for the elements. This conception seems similar to Peirce's *diagram* as explicated in (Stjernfelt 2011); under that analysis, a diagram is a concept useful for reasoning. In cognitive science, education, and computer science, that concept is split into a mental representation in the mind and external representation as arrangements in the world, props on a table, diagrams on a page, or gestures on a virtual page.

Not surprisingly, as both are created by minds, representations on the page bear some (structural) similarities (Shepard's (1984) isomorphisms) to mental representations (of course there are many differences as well). On the page, elements can be represented richly as depictions (cf. Kosslyn 1996; Shepard 1984) or minimally as labels or dots. Relations among elements can be represented spatially in a range of ways, by distance, grouping, ordering, place on page, by adding simple geometric forms that carry meaning, notably, lines, boxes, circles, and arrows. Many of these uses have been established empirically, in particular, in experiments that show that production of meanings match comprehension of the same meanings, a research program that can be regarded as empirical semiotics. In diagrams, for example, across a range of content, lines are comprehended and produced as relations (Kessell & Tversky 2011; Tversky & Lee 1999; Zacks & Tversky 1999), arrows as asymmetric relations (Heiser & Tversky 2006), and boxes as containers or sets (Kessell & Tversky 2011; Zacks & Tversky 1999).

The same visual/action/spatial forms that are drawn on the page are "drawn" in the air in gesture to represent the same meanings (e.g., Tversky et al. 2009). In gesture, elements can be established by using points or fists or tilts of the head or the eyes to point or by depictive gestures, sometimes accompanied by names. Relations among elements can be established by proximity of hands or head nods that refer to the elements, by lines, boxes, circles, and arrows conveyed by the hands. For Glasses, the glasses were often indicated by two groups of three points each, and they were grouped by proximity in space or time or both into those that were empty and those that were full. The action of pouring was typically indicated by a curved arrow "drawn" by a finger in the air or by pen on the page from the middle of the group of full glasses to the middle of the group of empty ones. For the environments, the locations of landmarks were indicated by points, and the paths between them by lines made by fingers or "chops" made by the side of the hand, like dots or blobs for locations and lines for paths in maps. Points or dots and lines by pen or finger form the backbone for networks, and networks can represent an enormous range of ideas and the relations among them. These simple forms, dots, lines, swipes, frames, arrows, and more, whether drawn or gestured, carry general meanings that are specified in context. The meanings are shared, produced by some and comprehended the same way by others (e.g., Tversky et al. 2000). They constitute a rudimentary vocabulary, a small set of semantic elements, often domain-specific, that can be combined to represent and convey the characteristics, relations, structure, behavior, and action of a large range of concepts.

Of course gesture and sketching differ in important ways. Gestures are actions, and fleeting. Sketches are static, and stable. "Sketch" here stands for marks on a page but also arrangements of props in the world. Sketches can be readily reexamined and revised, an advantage, especially for inference, discovery, design, and creativity. Gestures are fleeting, so although they are used to create representations, those representations must be kept in mind. Because sketches are static, conveying action and change in sketches and diagrams can be challenging. Inferring action, change, behavior, process, and causality from sketches can be difficult, especially for novices. Arrows help, but can be ambiguous (Heiser and Tversky, 2006). Animations, because they often do not segment action into natural units, because they show but do not explain, and because they are typically too complex and fast to grasp, rarely help (e.g., Tversky 2011; Tversky et al. 2002). Because gestures are actions, they seem to be especially good at conveying action (e.g., Cartmill et al. 2010; Jamalian & Tversky 2012; Kang et al. 2012). However, although gestures are often actions in form, they do not affect things in the world; they affect things in mind. They are not actions in and of themselves, rather, they are representations of mental or physical actions. Because they do not affect things in the world, their communicative intent can be readily discerned. Gesture and sketch are intrinsically related, sometimes even interchangeable; sketches are created by actions of the hands, and the actions that make marks on the page are referred to as gestures in the communities that study sketching. Gesture and sketching are often integrated in communication, complementing and supplementing one another. In explaining complex systems to others, people use gesture, diagram, and sketch (Engle 1998; Kang et al. 2015). They use gestures on sketches to refer to particular entities or parts of entities, they use gesture to "animate" diagrams by depicting actions or changes, they use gestures to create virtual diagrams, and then gesture on the virtual diagrams in the same ways that they gesture on actual diagrams (Kang et al. 2015). Thus, both sketches and gesture can establish representations and convey actions, but sketches have an advantage for representations and gestures for action.

Representations can be created in the mind as well as in the world. Importantly, we have seen that creating representations in the world helps to create them in the mind. Representations can also be created by language, but the relations between language and the represented world are arbitrary, and the relations between visual communication and the represented world are not. This is not to discount the role of language; on the contrary, it highlights the enormous power of language to create meaning. Canonical face-to-face communication is a harmonious blend of multi-media: language, gesture, intonation, props, and the world, complementing and supplementing each other to create meaning (e.g., Clark 1996). Here, we have shown that thinking, communicating with one's self, is often the same.

7 Where Does Thinking Come From?

The contents of thought have been regarded as internalized perceptions (e.g., Shepard 1984) and the actions of thought as internalized actions (cf. Bruner 1966; Vygotsky 1962), a view mirrored in the ways we talk about thinking. Among other things, we collect, sort, arrange, and discard our thoughts and ideas. The claim that actions of thought are like actions of the body receives support from accumulating brain research showing activation in motor or premotor cortices from diverse mental actions, for example, putting into memory, mental rotation, and counting (e.g., Andres et al. 2007; Braver et al. 1997; Eisenegger et al. 2007; Ganis et al. 2000; Kansaku et al. 2007; Manoach et al. 1997). That the contents of thought and actions of thought are internalized encourages mental bricolage, allowing the mind to decompose them and recompose them, to create new contents of thought and new actions on thought from parts of old ones. Examples of this are numerous, from the everyday creativity that allows us to replace a missing button with a safety pin to the educated creativity underlying inventing new theories and new elements.

If the contents of thought, mental representations, are at least in part internalized perceptions and the actions of thought are at least in part internalized actions, then reexternalizing thought as sketches and gestures should augment thinking by extending the mind. Evidence abounds that reexternalizing the contents of thought as arrangements of things in the world helps not only thought but memory and performance as well; paper and pencil, or its equivalents, for navigation, math, design, operating or assembling equipment, understanding or discovery in science, and more. Evidence that and how thinking is augmented by reexternalizing the thinking as gestures, as actions in but not on the world, has been slower to recognize and to accumulate.

8 *Spraction*: Integrating Space, Action, and Abstraction

As has been seen, both gesture and sketch can be used quite directly to represent information, concrete and abstract by mapping elements and relations of the information to marks in space and relations in space, on paper or in air. Gestures are typically distinguished from actions: gestures are actions that affect or create meaning whereas actions affect or create things in the world. Yet, they are intrinsically intertwined, as actions create sketches that affect or create meaning. Actions arrange things in the world that affect or create meaning. That is, gestures and sketches design thought. The world designed by human actions, of guests around a dinner table, of books on shelves, of buildings and streets in cities affect and create meaning. Those actions that arrange things in space, that design space, form regular patterns in space, and those patterns create meaning. Lines linking buildings on streets and seats in auditoria, boxes enclosing families in buildings and clothing in drawers, orders in bookshelves, hierarchies in piles of small and large plates and bowls in cupboards, one-to-one correspondences in table settings, repetitions and symmetries in buildings, repetitions in routines in time. These organizations signify: they instantiate and represent relations and orders and dimensions and categories and hierarchies of categories and symmetries and repetitions and one-to-one correspondences. These patterns, created by actions, form good Gestalts, and catch the eye. Because they are highly organized, they are likely to be interpreted as designed, not given in the natural world. Because they are viewed as designed, they invite interpreting the intentions and abstractions underlying the design. The actions that design them, actions on real objects, putting, collecting, lifting, sliding, inserting, piling, pushing, pulling, and more, are incorporated into gestures that represent mental actions, actions on objects of thought. The patterns created by actions in space, lines, boxes, repetitions, symmetries, and more, are incorporated into diagrams that represent myriad organizations and relations. That is, the actions get coopted as gestures and the patterns get coopted into diagrams that represent the abstractions, creating an interlinked and interacting cycle of action, space, and abstraction, gesture, sketch, and meaning, that can be entered anywhere and has been dubbed “spraction” Tversky (2011). The designed world is replete with meaning.

Acknowledgements

BT is indebted to her collaborators on the other research surveyed, Azi Jamalian, Valeria Giardino, Seokmin Kang, Andrea Kantrowitz, Masaki Suwa, Julie Heiser, Paul Lee, and Jeff Zacks. We are grateful to Riccardo Fusaroli and two anonymous reviewers for helpful comments, and to the following grants for facilitating the research and/or preparing the manuscript: National Science Foundation HHC 0905417, IIS-0725223, IIS-0855995, and REC 0440103, and Office of Naval Research NOOO14-PP-1-O649, N000140110717, and N000140210534.

Bibliography

Andres, M., X. Seron & E. Olivier. 2007. Contribution of Hand Motor Circuits to Counting. *Journal of Cognitive Neuroscience* 19. 563–576. https://doi.org/10.1162/jocn.2007.19.4.563.

Bilda, Z. & J. Gero. 2006. Reasoning with Internal and External Representations: A Case Study with Expert Architects. In R. Sun & N. Miyake (eds.), *Proceedings of the Cognitive Science Society*, 1020–1026. Mahwah, NJ: Lawrence Erlbaum.

Braver, T.S., J. Cohen, L. Nystrom, J. Jonides, E. Smith & D. Noll. 1997. A Parametric Study of Prefrontal Cortex Involvement in Human Working Memory. *Neuroimage* 1. 49–62. https://doi.org/10.1006/nimg.1996.0247.

Bruner, J.S. 1957. *Going Beyond the Information Given*. New York: Norton.

Bruner, J.S. 1966. On Cognitive Growth. In J. Bruner, R. Olver & P. Greenfield (eds.), *Studies in Cognitive Growth*, 1–29. Oxford: Wiley.

Carlson, R.A., M. Avraamides, M. Cary & S. Strasberg. 2007. What Do the Hands Externalize in Simple Arithmetic? *Journal of Experimental Psychology: Learning, Memory and Cognition* 33. 747–756. https://doi.org/10.1037/0278-7393.33.4.747.

Cartmill, E.A., S. Beilock & S. Goldin-Meadow. 2010. A Word in the Hand: Human Gesture Links Representations to Actions. *Philosophical Transactions of the Royal Society* B. 129–143. https://doi.org/10.1098/rstb.2011.0162.

Chu, M. & S. Kita. 2008. Spontaneous Gestures during Mental Rotation Tasks: Insights into the Microdevelopent of the Motor Strategy. *Journal of Experimental Psychology: General* 137. 706–723. https://doi.org/10.1047/a0013157.

Clark, H.H. 1996. *Using Language*. Cambridge: Cambridge University Press.

Donald, M. 1991. *Origins of the Modern Mind*. Cambridge: Harvard University Press.

Duncker, K. 1945. On Problem Solving. *Psychological Monographs* 68. 270.

Eisenegger, C., U. Herwig & L. Jäncke. 2007. The Involvement of Primary Motor Cortex in Mental Rotation Revealed by Transcranial Magnetic Stimulation. *Europen Journal of Neuroscience* 4. 1240–1244. https://doi.org/10.1111/j.1460-9568.2007.05354.x.

Emmorey, K., B. Tversky & H. Taylor. 2000. Using Space to Describe Space: Perspective in Speech, Sign, and Gesture. *Journal of Spatial Cognition and Computation* 2. 157–180. https://doi.org/10.1023/A:1013118114571.

Engle, R.A. 1998. Not Channels but Composite Signals: Speech, Gesture, Diagrams and Object Demonstrations Are Integrated in Multimodal Explanations. In M. Gernsbacher & S. Derry (eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Fitch, W.T., M. Hauser & N. Chomsky. 2005. The Evolution of the Language Faculty: Clarifications and Implications. *Cognition* 97. 179–210. https://doi.org/10.1016/j.cognition.2005.02.005.

Ganis, G., J. Keenan, S. Kosslyn & A. Pascual-Leone. 2000. Transcranial Magnetic Stimulation of Primary Motor Cortex Affects Mental Rotation. *Cerebral Cortext* 10. 175–180. https://doi.org/10.1093/cercor/10.2.175.

Gentner, D. & A. Stevens (eds.). 1983. *Mental Models*. Hillsdale, NJ: Erlbaum.

Goldin-Meadow, S. 2003. *Hearing Gesture: How Our Hands Help Us Think*. Cambridge: Belknap Press.

Goldin-Meadow, S., S. Cook & Z. Mitchell. 2009. Gesturing Gives Children New Ideas About Math. *Psychological Science* 20. 267–272. https://doi.org/10.111/j.1467-9280.2009.02297.x.

Gondry, M. 2013. *Is the Man Who Is Tall Happy? An Animated Conversation with Noam Chomsky*. Partizan Films.

Heiser, J. & B. Tversky. 2006. Arrows in Comprehending and Producing Mechanical Diagrams. *Cognitive Science* 30. 581–592. https://doi.org/10.1207/s15516709cog0000_70.

Heiser, J., B. Tversky & M. Silverman. 2004. Sketches for and from Collaboration. In J. Gero, B. Tversky & T. Knight (eds.), *Visual and Spatial Reasoning in Design III*, 68–78. Sydney: Key Centre for Design Research.

Hoeststetter, A.B. & M. Alibali. 2008. Visible Embodiment: Gestures as Simulated Action. *Psychonomic Bulletin and Review* 15. 495–514. https://doi.org/10.3758/PBR.15.3.495.

Holyoak, K.J. 1994. Problem Solving. In E. Smith & D. Osherson (eds.), *An Invitation to Cognitive Science: Thinking*, 267–296. Cambridge: MIT Press.

Hutchins, E. 1995. *Cognition in the Wild*. Cambridge: MIT Press.

Iverson, J.M. 1999. How to Get to the Cafeteria: Gesture and Speech in Blind and Sighted Children's Spatial Descriptions. *Developmental Psychology* 35. 1132–1142. https://doi.org/10.1037/0012-1649.35.4.1132.

Jamalian, A., V. Giardino & B. Tversky. 2013. Gestures for Thinking. In M. Knauff, M. Pauen, N. Sabaenz & I. Wachsmuth (eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Jamalian, A. & B. Tversky. 2012. Gestures Alter Thinking about Time. In N. Miyake, D. Peebles & R. Cooper (eds.), *Proceedings of the 35h Annual Conference of the Cognitive Science Society*, 551–557. Austin, TX: Cognitive Science Society.

Johnson-Laird, P. 1983. *Mental Models*. Cambridge: Harvard University Press.

Kang, S., B. Tversky & J. Black. 2012. From Hands to Minds: How Gestures Promote Action Understanding. In N. Miyake, D. Peebles & R. Cooper (eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, 551–557. Austin, TX: Cognitive Science Society.

Kang, S., B. Tversky & J. Black. 2015. Coordinating Gesture, Word, and Diagram: Explanations for Adult Experts and Young Novices. *Spatial Cognition and Computation* 15. 1–26. https://doi.org/10.1080/13875868.2014.958837.

Kansaku, K., B. Carver, A. Johnson, K. Matsuda, N. Sadato & M. Hallett. 2007. The Role of the Human Ventral Premotor Cortex in Counting Successive Stimuli. *Experimental Brain Research* 178. 339–350. https://doi.org/10.1007/s00221-006-0736-8.

Kantrowitz, A. 2014. *A Cognitive Ethnographic Study of Improvisational Drawing by Eight Contemporary Artists*: Columbia Teachers College dissertation.

Kendon, Adam. 2004. *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Kessell, A. & B. Tversky. 2006. Using Gestures and Diagrams to Think and Talk About Insight Problems. In R. Sun & N. Miyake (eds.), *Proceedings of the Meetings of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Kessell, A.M. & B. Tversky. 2011. Visualizing Space, Time, and Agents: Production, Performance, and Preference. *Cognitive Processing*.

Kintsch, W. 1996. *Comprehension: A Paradigm for Cognition*. Cambridge: Cambridge University Press.

Kirsh, D. 1995. The Intelligent Use of Space. *Artificial Intelligence* 73. 31–68. https://doi.org/10.1016/0004-3702(94)00017-U.

Kosslyn, S.M. 1996. *Image and Brain: The Resolution of the Imagery Debate*. Cambridge: MIT Press.

Krauss, R.M., Y. Chen & R. Gottesman. 2000. Lexical Gestures and Lexical Access: A Process Model. In *Language and Gesture*, 261–283. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511620850.017.

Maier, N.R.F. 1931. Reasoning in Humans II. The Solution of a Problem and its Appearance in Consciousness. *Journal of Comparative Psychology* 12. 181–194.

Manoach, D.S., G. Schlaug, B. Siewert, D. Darby, B. Bly, A. Benfield, R. Edelman & S. Warach. 1997. Prefrontal Cortex fMRI Signal Changes are Correlated with Working Memory Load. *Neuroreport* 8. 545–549. https://doi.org/10.1097/00001756-199701200-00033.

McNeill, D. 2005. *Gesture and Thought*. Chicago: Chicago University Press. https://doi.org/10.7208/chicago/9780226514642.001.0001.

McNeill, David. 1992. *Hand and Mind: What Gestures Reveal about Thought*. Chicago, IL: University of Chicago Press.

Netz, R. 1999. Linguistic Formulae as Cognitive Tools. *Pragmatics and Cognition* 7. 147–176. https://doi.org/10.1037/h0048495.

Newall, A., J. Shaw & H. Simon. 1958. Elements of a Theory of Human Problem Solving. *Psychological Review* 65. 151–166. https://doi.org/10.1037/h0048495.

Norman, D.A. 1993. *Things That Make Us Smart*. Boston: Addison-Wesley Longman.

Rauscher, F.H., R. Krauss & Y. Chen. 1996. Gesture, Speech, and Lexical Access: The Role of Lexical Movements in the Processing of Speech. *Psychological Science* 7. 226–231.

Schon, D.A. 1983. *The Reflective Practitioner*. New York: Harper Collins.

Schwartz, D.L. & J. Black. 1996. Shuttling between Depictive Models and Abstract Rules: Induction and Fallback. *Cognitive Science* 20. 457–497. https://doi.org/10.1207/s15516709cog2004_1.

Shepard, R.N. 1984. Ecological Constraints on Internal Representation: Resonant Kinematics of Perceiving, Imagining, Thinking, and Dreaming. *Psychological Review* 91. 417–447. https://doi.org/10.1037/0033-295X.91.4.417.

Stjernfelt, F. 2011. *Diagrammatology: An Investigation on the Borderlines of Phenomenology, Ontology, and Semantics*. New York: Springer.

Suchman, L. 1987. *Plans and Situated Actions*. Cambridge: Cambridge University Press.

Suwa, M. & B. Tversky. 1996. What Architects See in their Sketches: Implications for Design Tools. In *Human Factors in Computing Systems: Conference Companion*, 191–192. New York: ACM.

Suwa, M. & B. Tversky. 2003. Constructive Perception: A Skill for Coordinating Perception and Conception. In R. Alterman & D. Kirsh (eds.), *Proceedings of the Cognitive Science Society Meetings*. Mahwah, NJ: Erlbaum.

Suwa, M., B. Tversky, J. Gero & T. Purcell. 2001. Seeing into Sketches: Regrouping Parts Encourages New Interpretations. In J. Gero, B. Tversky & T. Purcell (eds.), *Visual and Spatial Reasoning in Design*, 207–2019. Sydney: Key Centre of Design Computing and Cognition.

Talmy, Leonard. 1983. How Language Structures Space. In H. Pick & L. Acredolo (eds.), *Spatial Orientation: Theory, Research, and Application*, 225–282. New York: Plenum Press.

Taylor, H.A. & B. Tversky. 1992. Spatial Mental Models Derived from Survey and Route Descriptions. *Journal of Memory and Language* 31. 261–282. https://doi.org/10.1016/0749-596X(92)90014-0.

Tversky, B., J. Heiser, P. Lee & M.-P. Daniel. 2009. Explanations in Gesture, Diagram, and Word. In K. Coventry, T. Tenbrink & J. A. Bateman (eds.), *Spatial Language and Dialogue*, 119–1131. Oxford: Oxford University Press. https://doi.org/10.1093/acprod:oso/9780199554201.003.0009.

Tversky, B. & P. Lee. 1999. Pictorial and Verbal Tool for Conveying Routes. In *Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science*, 51–64. Berlin: Springer.

Tversky, B., J. Morrison & M. Betrancourt. 2002. Animation: Can It Faciliate? *International Journal of Computer Studies* 57. 247–262. https://doi.org/10.1006/ijhc.2002.1017.

Tversky, B. & M. Suwa. 2009. Thinking With Sketches. In A. Markman (ed.), *Tools for Innovation*, Oxford: Oxford University Press.

Tversky, B., J. Zacks, L. P.U. & J. Heiser. 2000. Lines, Bloobs, Crosses, and Arrows: Diagrammatic Communication with Schematic Figures. In M. Anderson, P. Cheng & V. Haarslev (eds.), *Theory and Application of Diagrams*, Berlin: Springer. https://doi.org/10.1007/3-540-44590-0_21.

Tversky, Barbara. 2011. Visualizing Thought. *Topics in Cognitive Science* 3. 499–535.

Vygotsky, L. 1962. *Thought and Language*. Cambridge: MIT Press.

Wexler, M., S. Kosslyn & A. Berthoz. 1998. Motor Processes in Mental Rotation. *Cognition* 68. 77–94. https://doi.org/10.1016/S0010-0277(98)00032-8.

Wohlschläger, A. & A. Wohlschläger. 1998. Mental and Manual Rotatation. *Journal of Experimental Psychology: Human Perception and Performance* 24. 397–412.

Zacks, Jeffrey M. & B. Tversky. 1999. Bars and Lines: A Study of Graphic Communication. *Memory and Cognition* 27(6). 1073–1079.

Zhang, J-J. & D. Norman. 1994. Representations in Distributed Cognitive Tasks. *Cognitive Science* 18. 87–122.

Ralph Ewerth, Christian Otto, and Eric Müller-Budack

Computational Approaches for the Interpretation of Image-Text Relations

Abstract: In this paper, we present approaches that automatically estimate semantic relations between textual and (pictorial) visual information. We consider the interpretation of these relations as one of the key elements for empirical research on multimodal information. From a computational perspective, it is difficult to automatically "comprehend" the meaning of multimodal information and to interpret cross-modal semantic relations. One reason is that already the automatic understanding and interpretation of a single source of information (e.g., text, image, or audio) is difficult — and it is even more difficult to model and understand the interplay of two different modalities. While the complex interplay of visual and textual information has been investigated in communication sciences and linguistics for years, they have been rarely considered from a computer science perspective. To this end, we review the few currently existing approaches to automatically recognize semantic cross-modal relations. In previous work, we have suggested to model image-text relations along three main dimensions: cross-modal mutual information, semantic correlation, and the status relation. Using these dimensions, we characterized a set of eight image-text classes and showed their relations to existing taxonomies. Moreover, we have shown how the cross-modal mutual information can be further differentiated in order to measure image-text consistency in news at the entity level of persons, locations, and scene context. Experimental results demonstrate the feasibility of the approaches.

Keywords: multimodal semiotic analysis, semantic image-text classes, deep learning, multimodal news analytics, multimodal information retrieval, computer vision

1 Introduction

Multimodal communication is omnipresent (Bateman et al. 2017) and it helps convey information to recipients more effectively and efficiently, 1.) in various domains and genres, such as education, science, entertainment, news, and information, and 2.) in different types of media like newspapers, graphic novels, textbooks, television, video. When used jointly, two modalities normally yield additional information or even a new level of meaning. In many situations, a certain modality allows us to convey information in a way that is sometimes impossible when using another modality. For instance, it is not possible to exactly denote the date of birth

https://doi.org/10.1515/9783110725001-005

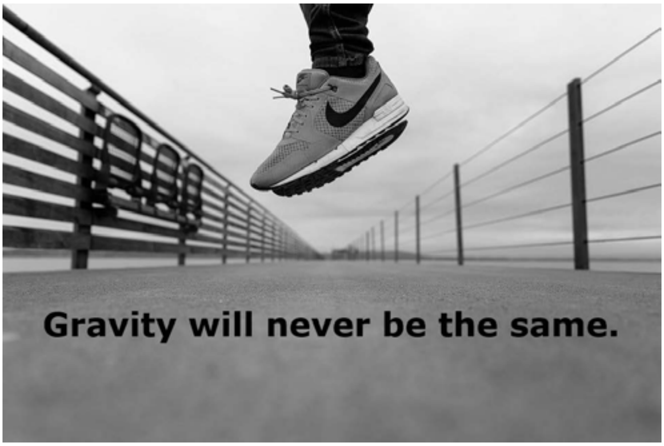


Fig. 1: An example of an image-text pair that illustrates the interplay between text and image content (PD by https://pixabay.com/service/license/).

in an image without using any textual information; on the other hand, it is not possible to exactly describe a human face or a plant's shape using textual information (without using an image or mathematical formulas) in a way such that a reader could exactly (re-)draw this face or plant (Henning & Ewerth 2017, 2018). In summary, information from different modalities is typically not completely identical, but especially their combination enables more effective communication (Bateman et al. 2017). This important property of multimodal information, or its inherent 'dissonance', is also present in multimedia content, e.g. between spoken text and scene content in videos.

The complex interplay of image and text, e.g., aspects of interdependence, consistency, or dissonance, has been researched in linguistics and communication sciences for quite some time; an overview is presented by Bateman (2014). One key to understand multimodal information is offered by semantic cross-modal references and relations. Such relations can exist on different semantic levels and in many forms: between image and text, video and speech, video and sound, etc. To date, no computational approaches exist that can interpret the interplay of multimodal information as given, for example, in the advertisement depicted in Figure 1.

Smeulders et al. (2000) defined the 'semantic gap' as "the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation", and identified this gap as one of the key challenges in image and multimedia retrieval research. Twenty years ago, the challenge was that information extraction from images was

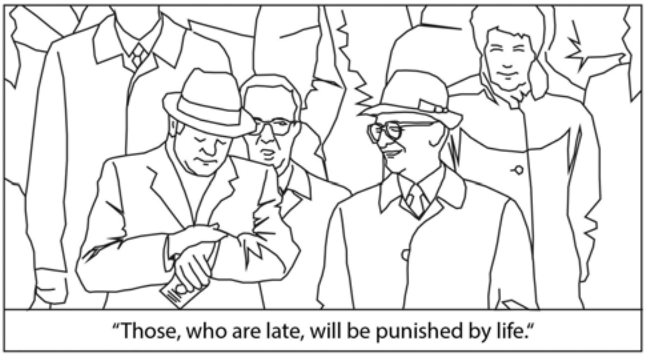


Fig. 2: An example of an image-text pair that requires historical and contextual knowledge. Without this knowledge it is not possible to interpret the photo as a key visual for Gorbachev's famous quote denoted below the image. (Rendering based on: https://www.welt.de/politik/20-jahre-mauerfall/article4872216/Als-Honecker-vor-20-Jahren-zuruecktreten-musste.html, last accessed: 07 December 2020.)

limited to low-level features by means of color, texture, and shape. Thus, multimedia and computer vision approaches aimed to solve the (perceptual) problem of object and scene recognition, considering visual concepts as semantic, high-level features. In fact, impressive progress has been reported for tasks such as object and visual concept recognition (Krizhevsky et al. 2012; He et al. 2016) or image captioning (Karpathy & Li 2015; Anderson et al. 2018) in recent years. However, these approaches aim to describe visual content focusing on objects, persons, etc., but lack capabilities of human scene interpretation going beyond the visible scene content, i.e., interpreting gestures, symbols, and other contextual information.

From a computational perspective, we identified a related issue as the 'information gap' between text and image data (Henning & Ewerth 2017, 2018). We argued that computer vision approaches work on a descriptive level, and thus often provide only superficial information and lack interpretation capabilities. Particularly, this is also the case for multimodal data. In this regard, neither the semantic nor the information gap have been closed by multimedia research yet, particularly not for multimodal data. For instance, to date no computational approaches exist that can interpret the image-text pair in Figure 2 in a historical context.

In this paper, we present two computational approaches that we have developed in previous work to classify and quantify image-text relations. The first approach defines a number of computable dimensions that capture different types of relations between image and text: *cross-modal mutual information*, *semantic correlation*, and the *status relation*. In Section 3, we describe these dimensions and show how meaningful image-text classes can be derived from them (Otto et al. 2019b, 2020). Further, we report experimental results for (automatic) classification using a deep learning approach. While the image-text dimensions are rather general, we differentiate the concept of cross-modal mutual information for the news domain in Section 4 (Müller-Budack et al. 2020). We briefly describe an approach to estimate the consistency between textual information and photo content in news articles. Experimental results demonstrate that the system can distinguish between correct and tampered cross-modal relations. This system is suitable to support empirical studies on large news corpora, or support users and journalists to find hints for false news.

2 Literature Review and Theoretical Background

2.1 Pattern Recognition for Vision and Language

We briefly review related work from computer science that focuses on visual information in images or on text information. In this context, many research efforts have addressed the automatic description of content. Machine learning (ML) and deep learning have contributed considerably to progress in related computer vision tasks, such as object recognition in images (Krizhevsky et al. 2012; Huang et al. 2017; Zoph et al. 2018), face recognition (Taigman et al. 2014; Schroff et al. 2015; Ding & Tao 2018; Deng et al. 2019), or geolocation estimation of photos (Weyand et al. 2016; Müller-Budack et al. 2018). Due to advances in neural networks (He et al. 2016; Vaswani et al. 2017; Kato et al. 2015), ever larger public datasets (Deng et al. 2009; Lin et al. 2014a; Benenson et al. 2019; Zhou et al. 2018a), and the development of better computing resources, current approaches can even achieve human-like performance in some tasks (Ewerth et al. 2017). Similar achievements can be reported for speech recognition (Chorowski et al. 2015; Amodei et al. 2016) and text analysis using models like Word2Vec (Mikolov et al. 2013) and BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2019). Deep learning has entailed substantial progress also in automatic image captioning. Following an encoder-decoder scheme, state-of-the-art approaches translate image features encoded with convolutional neural networks (CNNs) through language models that are based on recurrent neural networks (RNN) to descriptions, mainly using attention mechanisms (Vinyals et al. 2015; Xu et al. 2015; Anderson et al. 2018). However, these approaches are specialized for a single modality and mainly focus on the description of visual content rather than on interpretation.

Approaches for the analysis of multimodal information have also focused on recognizing or describing content, as the comprehensive survey on multimodal machine learning of Baltrusaitis et al. (2019) reveals. The survey summarizes the following tasks as typical multimodal analysis applications: audiovisual speech recognition (e.g., Ngiam et al. 2011; Afouras et al. 2018), event detection in videos (e.g., Lan et al. 2014), audiovisual emotion recognition (e.g., Kahou et al. 2016), media description (e.g., for images Karpathy & Li 2015; Anderson et al. 2018 or for video Zhou et al. 2018b; Aafaq et al. 2019), video summarization (e.g., Zhang et al. 2018a), and cross-modal retrieval (e.g., Zhen et al. 2019).

2.2 Linguistics and Communication Sciences

One direction of research in recent decades has dealt with the assignment of image-text pairs to distinct image-text classes. In pioneering work, Barthes (1977) discuss the respective roles and functions of text and images. He proposes a first taxonomy, introducing different types of (hierarchical) *status* relations between the modalities. If *status* is unequal, the classes *Illustration* and *Anchorage* are distinguished, otherwise their relation is denoted as *Relay*. Martinec & Salway (2005) extend Barthes' taxonomy and further divide the image-text combinations of *equal* rank into a *Complementary* and *Independent* class, indicating that the information content is either intertwined or equivalent in both modalities. They combine the taxonomy with the logico-semantics relations of Halliday & Matthiessen (2013), which were originally developed to characterize combinations of grammatical clauses. Martinec and Salway revised these grammatical categories to capture the specific logical relationships between text and image regardless of their *status*. In a very different tradition, McCloud (1993) focuses on comic books, whose characteristic is that image and text typically do not share information by means of depicted or mentioned concepts, albeit they have a strong semantic connection. He further denotes this relationship as *Interdependent* and argues that "pictures and words go hand in hand to convey an idea that neither could convey alone". Other authors mention the case of negative correlations between the mentioned or visually depicted concepts (for instance, Nöth 1995 and van Leeuwen 2005), denoting them as *Contradiction* or *Contrast*, respectively. Van Leeuwen states that they can be used intentionally, e.g., in magazine advertisements by choosing opposite colors or other formal features to draw attention to certain objects.

2.3 Computer Science Perspectives on Image-Text Relations

Although the semantic gap was recognized in the field of computer science and multimedia already twenty years ago (Smeulders et al. 2000), relations between image and text have been typically investigated at the level of image descriptions, neglecting (human-like) interpretation and deeper meanings of visual or multimodal information. We have suggested two metrics to characterize image-text relations: *cross-modal mutual information* and *semantic correlation* (Henning & Ewerth 2017, 2018), whose definitions will be described in detail in the next section. To the best of our knowledge, this has been one of the first attempts to model image-text relations in a more differentiated manner. Technically, an autoencoder with multimodal embeddings was proposed to learn these relations. The autoencoder consists of an encoder-decoder architecture, and its goal is to reduce the usually high-dimensional input data (e.g., 90 000 input neurons for a $300x300$ pixel image, in addition to text) to a compact, low-dimensional representation (around 2 000) that disregards redundant, uninteresting parts of the input data. While the encoder part transforms the input data to a low-dimensional representation, the decoder part learns to reconstruct the original input, i.e., it is forced to 'remember' only the salient parts of image and text. Since training does not require human annotations and learning is based on comparing input and output instead, this method can be used in a pre-training step when only a limited amount of labeled data is available to train the actual classifier.

Quite recently, there have also been other approaches that address image-text relations in a more differentiated manner. Zhang et al. (2018b) investigate image-text relations in advertisements and distinguish, for instance, between equivalent parallel and non-equivalent parallel information transfer. This work is further extended by Ye et al. (2019) in the ADVISE (Ads Visual Semantic Embedding) model, which interprets the rhetoric of visual advertisements using cross-modal embeddings and image embeddings for symbol regions. However, they do not regard previous work, e.g., from the field of linguistics: instead of discussing existing definitions (see Section 3) they define their own set of relations. Kruk et al. (2019), on the other hand, utilize the relation taxonomy of Marsh & White (2003) to model the author's intent in combining text and image in Instagram posts. Two kinds of image-text relations are suggested: the *contextual relation* between the literal meanings of the image and caption, and the *semiotic relationship* between the image and the caption.

3 Computable Image-Text Classes

In this section, we describe our computational model for image-text relations that consists of three dimensions. The model extends our initial set of the two dimensions *cross-modal mutual information* (CMI) and *semantic correlation* (SC) (Henning & Ewerth 2017, 2018) with the *status relation* (STAT), which is described in Section 3.1. Using these three dimensions, we can derive a set of computable image-text classes (Section 3.2; Otto et al. 2019b,a). The three image-text dimensions eventually yield eight image-text classes, which are partially compliant to classes in existing taxonomies: in particular, *Anchorage*, *Complementary*, *Illustration*, *Interdependent*, and *Uncorrelated*. Further, there are three classes for contrasting or contradicting (cross-modal) information: *Bad Anchorage*, *Contrasting*, and *Bad Illustration*.

As a point of departure, Martinec and Salway's taxonomy yields the classes *Illustration*, *Anchorage*, *Complementary*, and *Independent*. We disregard the class *Independent* since it is very uncommon that both modalities describe exactly the same information. We introduce the class *Interdependent* suggested by McCloud (1993), which in contrast to *Complementary* consists of image-text pairs where the intended meaning cannot be gathered from either text or image exclusively. While several categorizations do not consider negative semantic correlations at all, Nöth (1995), van Leeuwen (2005) and Henning & Ewerth (2017) consider this aspect as well. We believe that it is important to consider negative correlations, for instance, in order to identify less useful multimodal information, contradictions, mistakes, etc. Consequently, the classes *Contrasting*, *Bad Illustration*, and *Bad Anchorage* provide the negative counterparts for *Complementary*, *Illustration*, and *Anchorage*, respectively. Finally, we consider the case when text and image are *uncorrelated*.

While one objective of our work is to derive meaningful, distinctive, and comprehensible image-text classes, another contribution is their systematic characterization. For this purpose, the metrics *cross-modal mutual information* (CMI) and *semantic correlation* (SC) (Henning & Ewerth 2017, 2018) are not sufficient to model a wide range of image-text classes. It is apparent that the *status* relation, originally introduced by Barthes (1977), has been adopted by the majority of taxonomies established in the past (e.g., (e.g., Martinec & Salway 2005; Unsworth 2007), suggesting that this relation is essential to describe an image-text pair. As characterized by Barthes, the *status relation* describes how two modalities can hierarchically relate to one another reflecting their relative importance. Either the text supports the image (*Anchorage*), or the image supports the text (*Illustration*), or both modalities contribute equally to the overall meaning (e.g., *Complementary*). This encouraged us to extend the two-dimensional feature space of CMI and

SC with the *status relation* (*STAT*). In the next section, we provide definitions for the three metrics and subsequently infer a categorization of semantic image-text classes from them. The goal is to reformulate and clarify the interrelations between visual and textual content in order to make them applicable for computational approaches.

3.1 Dimensions of Image-Text Relations

Concepts and Entities

The following definitions are related to concepts and entities in images and text. Generally, many concepts and entities can be found in images ranging from the main focus of interest (e.g., a person, a certain object, a location) to barely visible or background details (e.g., a leaf of grass, a bird in the sky). Normally, the meaning of an image is related to the main objects in the foreground. When assessing relevant information in images, it is reasonable to regard these concepts and entities, which, however, adds a certain level of subjectivity in some cases. But most of the time the important entities can be determined.

Cross-Modal Mutual Information (CMI)

Depending on the (fraction of) mutual presence of concepts and entities in both image and text, the *cross-modal mutual information* ranges from 0 (no overlap of depicted concepts) to 1 (concepts in image and text overlap entirely), or mathematically $\mathrm{CMI} \in [0, 1]$. It is important to point out that CMI ignores any deeper semantic meaning, in contrast to *semantic correlation*. If, for example, a small man with a blue shirt is shown in the image, while the text talks about a tall man with a red sweater, the CMI would still be positive due to the mutual concept 'man'. But since the description is confusing and hinders interpretation of the multimodal information, the semantic correlation (SC, see below) of this image-text pair would be negative. Image-text pairs with high CMI can be found in image captioning datasets, for instance. The images and their corresponding captions often have a descriptive nature, which is why they have explicit representations in both modalities. In contrast, news articles or advertisements often have a loose connection to their associated images through mutual entities or concepts.

Semantic Correlation (SC)

The interpretation of the *semantic correlation* of image and text can range from coherent ($SC = 1$), over uncorrelated ($\mathrm{SC} = 0$) to contradictory ($\mathrm{SC} = -1$). This refers to concepts, descriptions and interpretation of symbols, metaphors, as well as to their relations to one another. Typically, an interpretation requires contextual information, knowledge, or experience and it cannot be derived exclusively from the entities in the text and the objects depicted in the image. The range of possible values is $[-1, 1]$, where a negative value indicates that the co-occurrence of an image and a text is contradicting and disturbs the comprehension of the multimodal content. This is the case if a text refers to an object in an image and cannot be found there, or has different attributes as described in the text. An observer might notice a contradiction and ask herself 'Do image and text belong together at all, or were they placed jointly by mistake?'. Conversely, a positive score suggests that both modalities share a semantic context or interpretation. The third possible option is that there is no semantic correlation between entities in the image and the text, yielding SC = 0.

Status Relation (STAT)

As introduced by Barthes (1977), *status* describes the hierarchical relation between an image and text with respect to their relative importance. Either the image is 'subordinate to the text' (STAT = T), implying an exchangeable image which plays the minor role in conveying the overall message of the image-text pair, or the text is 'subordinate to the image' (STAT = I), usually characterizing text with additional information (e.g., a caption) for an image that is the center of attention. An *equal status* (STAT = 0) describes the situation where image and text are equally important to convey the overall message. Images which are 'subordinate to text' (class *Illustration*) 'elucidate' or 'realize' the text. This is the case if a text describes a general concept and the associated image shows a specific example of that concept. Examples for the class *Illustration* can be found in textbooks and encyclopedias. Conversely, in the class *Anchorage* the text is 'subordinate to the image'. This is the case if the text answers the question 'What can be seen in this image?'. It is common that (implicit and explicit) direct references to objects in the image can be found and the readers are informed what they are looking at. This type of image-text pair can be found in newspapers or scientific documents, but also in image captioning data sets. The third possible state of a *status relation* is 'equal', which describes an image-text pair where both modalities contribute individually to the conveyed information or either part contains some details that the other one does not. According to Barthes (1977), this class describes the situation where the information depicted in either modality is part of a more general message and together they elucidate information on a higher level that neither could do alone.

3.2 Defining Classes of Image-Text Relations

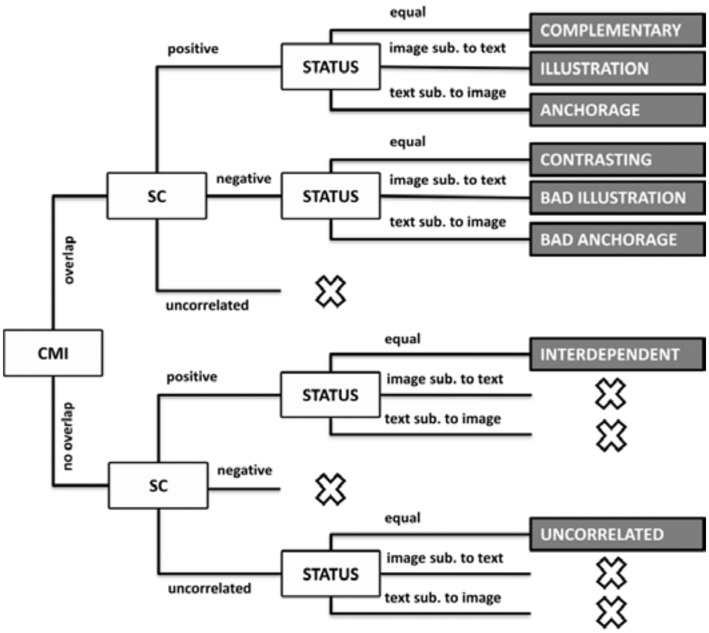


Fig. 3: Our categorization of image-text relations. Discarded subtrees or leaves are marked by an X for clarity; no hierarchical relations are implied.

In this section, we show how the combination of our three metrics can be mapped to distinctive image-text classes (see also Figure 3). For this purpose, we simplify the data value space for each dimension. The level of semantic correlation can be represented by the interval $[-1, 1]$. In our first proposal for CMI and SC, we suggested five (ordinal) levels of CMI and SC (Henning & Ewerth 2017, 2018). Here, we omit these intermediate levels since the general idea of positive, negative, and uncorrelated image-text pairs is sufficient for the task of assigning image-text pairs to distinct classes. Therefore, the possible states of semantic correlation (SC) are: SC $\in \{-1, 0, 1\}$. For a similar reason, finer levels for CMI are omitted, resulting in two possible states for CMI $\in \{0, 1\}$, which correspond to *no overlap* and *overlap*. Possible states of *status relations* are STAT $\in \{T, 0, I\}$: *image subordinate to text* (STAT $= T$), *equal status* (STAT $= 0$), and *text subordinate to image* (STAT $= I$). If approached naively, there are then $2 \times 3 \times 3 = 18$ possible combinations of SC, CMI, and STAT. A closer inspection reveals that (only) eight of these classes match with existing taxonomies in communication sciences, however. The remaining ten classes can be discarded since they cannot occur in practice or do not make sense. The reasoning behind this is given below after we have defined the eight classes that form the categorization.

Uncorrelated ($CMI = 0, SC = 0, STAT = 0$)

This class contains image-text pairs that do not belong together in an obvious way. They neither share entities and concepts nor is there an interpretation for a semantic correlation (e.g., see Figure 4, left).

Complementary ($CMI = 1, SC = 1, STAT = 0$)

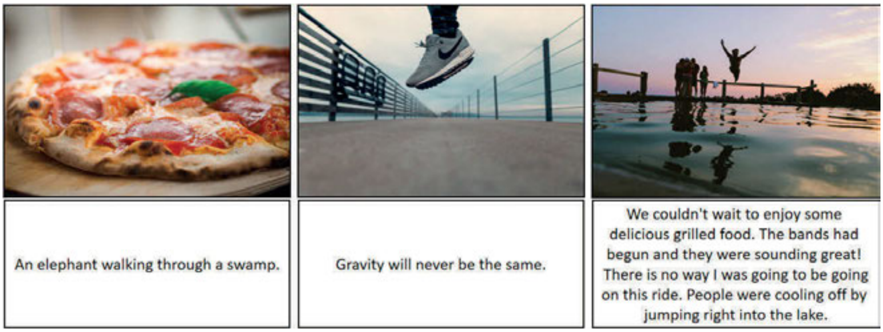
The class *Complementary* comprises the classic interplay between visual and textual information, i.e., both modalities share information but also provide information that the other one does not. Neither of them is dependent on the other and their status relation is equal. It is important to note that the amount of information is not necessarily the same in both modalities. The most significant factor is that an observer is still able to understand the key information provided by either of the modalities alone (Figure 4, right). The definitions of the next two classes clarify this further.

Interdependent ($CMI = 0, SC = 1, STAT = 0$)

This class includes image-text pairs that do not share entities or concepts by means of mutual information, but are related by a semantic context. As a result, their combination conveys a new meaning or interpretation which neither of the modalities could have achieved on its own. Such image-text pairs are prevalent in advertisements where companies combine eye-catching images with funny slogans supported by metaphors or puns, and without actually naming their product (Figure 4, middle). Another genre that relies heavily on these *interdependent* examples are comics or graphic novels, where speech bubbles and accompanying drawings are used to tell a story. Interdependent information is also prevalent in movies and TV material.

Anchorage ($CMI = 1, SC = 1, STAT = I$)

In contrast, the *Anchorage* class is an image description that acts as a supplement for an image. Barthes states that the role of the text in this class is to fix the interpretation of the visual information as intended by the author of the image-text pair (Barthes (1977)). It answers the question ‘What is it?’ in a more or less detailed manner. This is often necessary since the possible meanings or interpretations of



An elephant walking through a swamp.

Gravity will never be the same.

We couldn't wait to enjoy some delicious grilled food. The bands had begun and they were sounding great! There is no way I was going to be going on this ride. People were cooling off by jumping right into the lake.

Fig. 4: Examples of the *Uncorrelated* (left), *Interdependent* (middle) and *Complementary* (right) classes. All images Ⓟ by https://pixabay.com/service/license/.

an image can noticeably vary and the caption is provided to pinpoint the author's intention. Therefore, an *Anchorage* can be a simple image caption, but also a longer text that elucidates the hidden meaning of a painting. It is similar to *Complementary*, but the main difference is that in *Anchorage* the text is subordinate to image (see Figure 5).

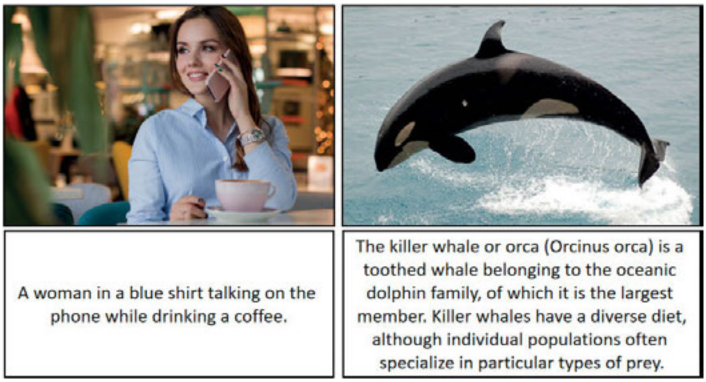
Illustration ($CMI = 1, SC = 1, STAT = T$)

The class *Illustration* contains image-text pairs where the visual information is subordinate to the text and has therefore a lower *status*. An instance of this class could be, for example, a text that describes a general concept and the accompanying image depicts a specific example (Figure 5). A distinctive feature of this class is that the image may be replaced by a very different image without rendering the constellation invalid. If the text is a definition of the term 'mammal', for example, it does not matter if the image shows an elephant, a mouse, or a dolphin. Each of these examples would be valid in this scenario. In general, the text is not dependent on the image to provide the intended information.

3.3 Impossible Image-Text Relations

Contrasting ($CMI = 1, SC = -1, STAT = 0$)
Bad Illustration ($CMI = 1, SC = -1, STAT = T$)
Bad Anchorage ($CMI = 1, SC = -1, STAT = I$)

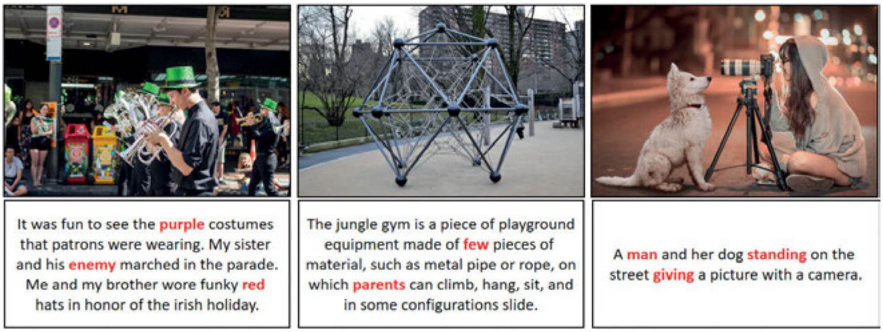
These three classes are the counterparts to *Complementary, Illustration*, and *Anchorage*: they share their primary features, but have a **negative SC** (see Figure 6). In other words, the transfer of knowledge is impaired due to inconsistencies or



A woman in a blue shirt talking on the phone while drinking a coffee.

The killer whale or orca (Orcinus orca) is a toothed whale belonging to the oceanic dolphin family, of which it is the largest member. Killer whales have a diverse diet, although individual populations often specialize in particular types of prey.

Fig. 5: Examples for *Anchorage* (left) and *Illustration* (right) classes. All images Ⓟ by https://pixabay.com/service/license/.



It was fun to see the purple costumes that patrons were wearing. My sister and his enemy marched in the parade. Me and my brother wore funky red hats in honor of the irish holiday.

The jungle gym is a piece of playground equipment made of few pieces of material, such as metal pipe or rope, on which parents can climb, hang, sit, and in some configurations slide.

A man and her dog standing on the street giving a picture with a camera.

Fig. 6: Examples for the *Contrasting* (left), *Bad Illustration* (middle), and *Bad Anchorage* (right) classes. All images Ⓟ by https://pixabay.com/service/license/.

contradictions when jointly viewing image and text (Henning & Ewerth 2017). In contrast to *uncorrelated* image-text pairs, these classes share information and obviously they belong together in a certain way, but particular details or characteristics are contradicting. For instance, a *Bad Illustration* pair could consist of a textual description of a bird, whose most prominent feature is its colorful plumage, but the bird in the image is actually a grey pigeon. This can be confusing and an observer might be unsure if she is looking at the right image. Similarly, contradicting textual counterparts exist for each of these classes. In Section 3.4, we describe how we generate training samples for these classes.

The eight classes described above form the categorization as shown in Figure 3. The following ten combinations of metrics were discarded, since they do not yield meaningful image-text pairs.

Cases A: $CMI = 0, SC = -1, STAT = T, 0, I$

These three classes cannot exist: If the shared information is zero, then there is nothing that can contradict. As soon as a textual description relates to a visual concept in the image, there is cross-modal mutual information and so $CMI > 0$.

Cases B: $CMI = 0, SC = 0, STAT = T, I$

The metric combination $CMI = 0, SC = 0, STAT = 0$ describes the class *Uncorrelated* of image-text pairs which are neither in contextual nor visual relation to one another. Since it is not intuitive that a text is subordinate to an uncorrelated image or vice versa, these two classes are discarded.

Cases C: $CMI = 0, SC = 1, STAT = T, I$

Image-text pairs in the class *Interdependent* $(CMI = 0, SC = 1, STAT = 0)$ are characterized by the fact, that, even though they do not share any information, they still complement each other by conveying additional or new meaning. Due to the nature of this class, a subordination of one modality to the other is not plausible: Neither of the conditions for the states *image subordinate to text* and *text subordinate to image* can be fulfilled due to lack of shared concepts and entities. Therefore, these two classes are discarded.

Cases D: $CMI = 1, SC = 0, STAT = T, 0, I$

As soon as there is an overlap of essential depicted concepts there has to be a minimum of semantic overlap. We consider entities as essential if they contribute to the overall information or meaning of the image-text pair. This excludes trivial background information such as the type of hat a person wears in an audience behind a politician giving a speech. The semantic correlation can be minor, but it would still correspond to $SC = 1$ according to the definition above. Therefore, the combination $CMI = 1, SC = 0$ and the involved possible combinations of $STAT$ are discarded.

3.4 Datasets

In order to successfully train a CNN-based machine learning model, a large number of annotated training examples is required to improve the likelihood of the model to generalize for a certain task (e.g., the Russakovsky et al. 2015 dataset for concept classification). If not enough training examples are provided, the model might only memorize the few given examples and their solutions rather than learn what their similarities and distinct features are. Therefore, we utilized several datasets available on the Web that are related to image-text classes. For example, classes like *Complementary* or *Anchorage* are available from several sources and can therefore be scraped from websites utilizing automatic web crawling techniques. Other classes, like *Uncorrelated*, do not naturally occur on the Web, but can be generated with little effort. But classes like *Contrasting* or *Bad Anchorage* are rare. While they do exist and it is desirable to detect these image-text pairs as well, there is no abundant source of such examples that could be used to train a robust classifier.

Only a few datasets are publicly available that contain images and corresponding text in form of cohesive sentences, i.e., not simply based on tags and keywords. Two examples are the image captioning dataset MSCOCO (Lin et al. 2014b) and the Visual Storytelling dataset VIST (Huang et al. 2016). A large number of examples can be easily taken from these datasets, namely for the classes *Uncorrelated*, *Complementary*, and *Anchorage*. Specifically, the underlying hierarchy of MSCOCO is exploited to ensure that two randomly picked examples are not semantically related to one another; we then join the caption of one sample with the image of the other one to form *Uncorrelated* samples. In this way, we gathered 60 000 ***Uncorrelated*** training samples.

The VIST dataset has three types of captions for their five-image-stories. The first one 'Desc-in-Isolation' resembles the generic image-caption dataset and can be used to generate examples for the class ***Anchorage***. These short descriptions are similar to MSCOCO captions, but slightly longer, so we decided to use them. Around 62 000 examples were generated this way. The pairs represent this class well, since they include textual descriptions of the visually depicted concepts without any low-level visual concepts or added interpretations. The second type of VIST captions 'Story-in-Sequence' is used to create ***Complementary*** samples by concatenating the five captions of a story and pairing them randomly with one of the images of the same story. Using this procedure, we generated 33 088 examples.

While there are certainly many more possible constellations of *complementary* content from a variety of sources, the various types of stories of this dataset already offer a solid basis. The same argumentation holds for the ***Interdependent*** class. We had to manually annotate a set of about 1 007 entries of Hussain et al.'s Internet Advertisements data set (Hussain et al. 2017) to generate these image-text pairs. While they exhibit the right type of image-text relations, the accompanying slogans (in the image) are not annotated separately and optical character recognition did not achieve high accuracy due to ornate fonts, etc. Furthermore, some image-text pairs had to be removed, since some slogans specifically mention the product name. This contradicts the condition that there is no overlap between depicted concepts and textual description, i.e., $CMI = 0$.

The ***Illustration*** class was established by combining one random image for each concept of the ImageNet dataset Russakovsky et al. (2015) with the summary of the corresponding article of the English Wikipedia, in case it existed. This nicely fits the nature of the class since the Wikipedia summary often provides a definition including a short overview of a concept. An image of the ImageNet class with the same name as the article should be a replaceable example image of that concept.

The three remaining classes ***Contrasting, Bad Illustration*** and ***Bad Anchorage*** occur rarely, as mentioned before, and are hard to crawl automatically. Therefore, it is not possible to automatically crawl a sufficient quantity of samples, since there is no abundant source of image-text combinations that intentionally contradict one another. To circumvent this problem, we explored transforming the respective positive counterparts by replacing 530 keywords[1] (adjectives, directional words, colors) by antonyms and opposites in the textual description of the positive examples to make them less comprehensible. For instance, 'tall man standing in front of a green car' was transformed into a 'small woman standing behind a red car'. While this does not absolutely break the semantic connection between image and text, it surely describes certain attributes incorrectly and so impairs the accurate understanding and subsequently justifies the label of $SC = -1$. This strategy allowed us to transform a substantial number of the 'positive' image-text pairs into their negative counterparts.

Finally, for all classes we truncated the text if it exceeded 10 sentences. The reason is that we wanted to avoid the network associating certain image-text classes with specific text lengths, since this will not always be the case for samples outside our training data. For example, the abstracts gathered from Wikipedia were on average much longer than the image captions from the MSCOCO dataset.

In total, the final dataset consists of $224\,856$ image-text pairs. Table 1 gives an overview of the data distribution sorted by class. The imbalanced nature of the distribution (up to $1 : 62$), which is caused by the challenges gathering the respective samples, is reduced to around $1 : 16$ when looking at the dataset from the perspective of the three metrics in Table 2. In our experiments, we evaluate both the class label and the metric label.

[1] http://www.myenglishpages.com/site_php_files/vocabulary-lesson-opposites.php., last accessed: 11/23/2017

Tab. 1: Distribution of class labels in the generated dataset.

Class	Num. of Samples
Uncorrelated	60 000
Interdependent	1 007
Complementary	33 088
Illustration	5 447
Anchorage	62 637
Contrasting	31 368
Bad Illustration	4 099
Bad Anchorage	27 210

Tab. 2: Distribution of metric labels in the generated dataset.

Class	Num. of Samples
STAT T	125 463
STAT 0	9 546
STAT I	89 847
SC -1	62 677
SC 0	60 000
SC 1	102 179
CMI 0	61 007
CMI 1	163 849

3.5 Experimental Setup and Results

The dataset was split into a training set and a manually verified test set to ensure high-quality labels. It initially contained 800 image-text pairs, where for each of the eight classes 100 examples were taken out of the automatically crawled and augmented data. The remaining 239 307 examples were used to train four different models with the Python Deep Learning Library TensorFlow (Abadi et al. 2016): three models for a 'cascade' classifier and one for the 'classic' approach. The cascade approach consists of three classifiers for the dimensions CMI, SC, and STAT (Section 3.2), whereas in the 'classic' approach the neural network directly outputs the image-text class (one of eight). The details are described in the corresponding paper (Otto et al. 2020).

To assure highly accurate ground-truth data for our test set, we asked three persons of our group (one of them a co-author) to manually annotate the 800 image-text pairs. Each annotator received an instruction document that contained short definitions of the three metrics (Section 3.1), the categorization in Figure 3, and one example per image-text class (similar to Figures 4-6). The quality of a dataset can

Tab. 3: Comparison of the automatically generated labels with the annotations of the three volunteers (i.e., *ground-truth data*) and the resulting number of samples per class in the test set.

Class	**Uncorr.**	**Interdep.**	**Compl.**	**Illustration**
Precision	98.7%	96.3%	88.0%	80.7%
Recall	69.2%	97.6%	83.8%	83.7%
F1-Score	81.3%	97.0%	85.9%	82.2%
#Samples	149	100	106	95

Class	**Anchorage**	**Contrasting**	**Bad Illu.**	**Bad Anch.**
Precision	87.3%	78.3%	69.0%	87.0%
Recall	90.3%	89.0%	98.6%	91.9%
F1-Score	88.8%	83.3%	81.2%	89.4%
#Samples	95	87	71	95

be evaluated by computing the inter-coder agreement, measuring how much the annotators agreed on the labeling of the samples. We utilized Krippendorff's alpha (Krippendorff (1970)) which yielded a value of $\alpha = 0.847$ (across all annotators, samples, and classes). A class label was assigned to a sample, if the majority of annotators agreed on it. Besides the eight image-text classes, the annotators could also mark a sample as *Unsure* to denote that an assignment was not possible. If *Unsure* was the majority of votes, the sample was not considered for the test set. This only applied for two pairs, which reduced the size of the final test set to 798.

Comparing the human labels with the labels from crawled image-text pairs, i.e., based on the Web resource, allows us to evaluate the quality of the data acquisition process. Therefore we computed how well the automatic labels matched the human ground-truth labels (Table 3). The low recall for the class *Uncorrelated* indicates that there were uncorrelated samples in the other data sources that we exploited. The *Bad Illustration* class has the lowest precision and was mostly confused with *Illustration* and *Uncorrelated* (cf. Table 4), that is the human annotators considered the automatically 'augmented' samples either as still valid or uncorrelated.

The results for predicting image-text classes using both the 'classic approach' (Table 4) and the 'cascade approach' (Table 5) are presented in confusion matrices in terms of precision and recall. For a better comparison, Figure 7 shows the individual performance for each image-text class. The accuracy of the classifiers for CMI, SC, and STAT ranges from 83.8% (STAT) over 84.6% (SC) to 90.3% (CMI), while the two classification variations for the image-text classes achieved an accuracy of 74.3% (*cascade*) and 80.8% (*classic*). An advantage of the *cascade* approach is

Tab. 4: Confusion matrix for the “classic” classifier on the test set of 798 image-text pairs. The rows show the ground-truth, while the columns show the predicted samples. (The *Undefined* column was added for better comparability with Table 5.)

Class	Undef.	Uncorrelated	Interdep.	Compl.	Illustration	Anchorage	Contrasting	Bad Illust.	Bad Anch.	Sum
Uncorrelated	-	**67**	3	5	23	34	5	11	1	149
Interdependent	-	0	**94**	0	0	5	0	0	1	100
Complementary	-	0	0	**93**	0	4	9	0	0	106
Illustration	-	0	0	0	**84**	0	0	11	0	95
Anchorage	-	2	2	0	2	**83**	0	0	6	95
Contrasting	-	0	0	3	0	0	**84**	0	0	87
Bad Illustration	-	0	0	0	2	0	0	**69**	0	71
Bad Anchorage	-	2	0	0	0	21	1	0	**71**	95
Precision	-	94.4%	94.9%	92.1%	75.7%	56.5%	84.8%	75.8%	89.9%	-
Recall	-	45.0%	94.0%	87.7%	88.4%	87.4%	96.5%	97.2%	74.7%	-
F1-Score	-	60.9%	94.5%	89.9%	81.6%	68.6%	90.3%	85.2%	81.6%	-

Tab. 5: Confusion matrix for the ‘cascade’ classifier on the test set of 798 image-text pairs. The rows show the ground-truth, while the columns show the predicted samples.

Class	Undef.	Uncorrelated	Interdep.	Compl.	Illustration	Anchorage	Contrasting	Bad Illust.	Bad Anch.	Sum
Undefined	**0**	0	0	0	0	0	0	0	0	0
Uncorrelated	2	**96**	4	7	21	1	4	13	1	149
Interdependent	3	3	**92**	1	0	1	0	0	0	100
Complementary	1	0	1	**93**	0	2	9	0	0	106
Illustration	1	0	0	0	**82**	0	0	12	0	95
Anchorage	11	4	5	25	1	**41**	2	1	5	95
Contrasting	0	0	0	2	0	0	**85**	0	0	87
Bad Illustration	0	0	0	0	8	0	0	**63**	0	71
Bad Anchorage	9	2	0	4	0	6	33	0	**41**	95
Precision	-	91.4%	90.2%	70.5%	73.2%	80.4%	63.9%	70.8%	87.2%	-
Recall	-	64.4%	92.00%	87.7%	86.3%	43.2%	97.7%	88.7%	43.1%	-
F1-Score	-	75.6%	91.10%	78.2%	79.2%	56.2%	77.3%	78.8%	57.8%	-

Tab. 6: Performance of the single metric classifiers.

-	CMI 0	CMI 1	SC 0	SC 1	SC -1	STAT 0	STAT T	STAT I
Precision	87.7%	91.4%	81.8%	84.2%	86.6%	82.5%	82.2%	92.8%
Recall	80.3%	94.9%	90.5%	64.4%	88.4%	90.5%	100.0%	54.2%
F1-Score	83.9%	93.1%	85.9%	73.0%	87.5%	86.3%	90.2%	68.4%

the prediction of the 'impossible' combinations — as argued above — of CMI, SC, and STAT. In this regard, the low number of (false) predictions for the 'undefined' image-text class supports our arguments. The lower performance of the *cascade* approach is mainly related to the lower recall for the classes *Anchorage* and *Bad Anchorage*, which itself is caused by the low recall for predicting the status class *text is subordinate to image*. Overall, the *classic* approach clearly outperforms the *cascade* approach by 6.1% in terms of accuracy.

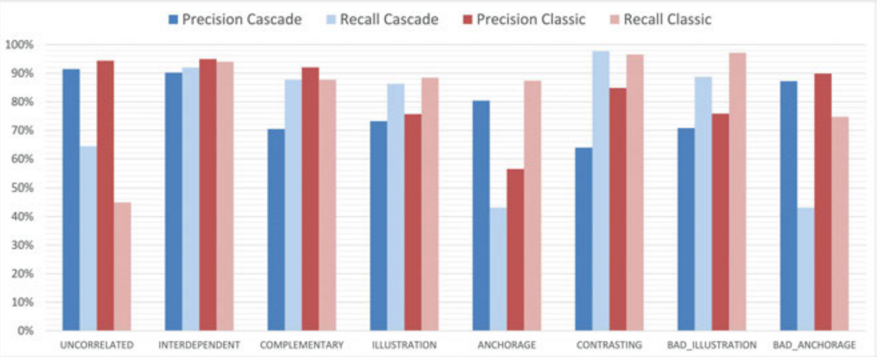


Fig. 7: Results for both classifiers.

4 Differentiating Cross-Modal Mutual Information in Multimodal News

We have further improved our model with a more fine-grained and differentiated model for cross-modal information. In this regard, we have suggested a system that automatically measures the consistency between pairs of image and text in news (Müller-Budack et al. 2020). The verification of consistency is realized through measures of cross-modal similarities for different entity types (persons, locations, and events) and a more general context. For a given image-text pair, we extract named entities from the text using named entity linking. Subsequently, we automatically download reference images from the Web, e.g., using image search engines like *Bing* or *Google*. These images serve as an input for the visual verification of the entities to the accompanying news image.

Visual features are obtained by appropriate state-of-the-art computer vision approaches, which are used in conjunction with measures of cross-modal similarity to quantify the cross-modal consistency. The following measures are suggested: *Cross-modal Person Similarity* (CMPS) and *Cross-modal Event Similarity* (CMES) *Cross-modal Location Similarity* (CMLS) and *Cross-modal Context Similarity* (CMCS).

We explain the approach for the task of *Cross-modal Person Similarity*, as illustrated in Figure 8. Given an image-text pair, we extract the entities from the text. Then, we gather a maximum of k example images using image search engines such as *Google* or *Bing* for each person $p \in P$ that was extracted from the named entity linking. Since these images can contain noise, i.e., showing other or additional persons, a filtering step is necessary. Thus, feature vectors are extracted for each face detected in the images. Since images that depict the same persons should have similar feature vectors, we first aim to assign them to groups (also called clusters). We assume that the majority cluster, i.e., the cluster with the maximum number of entries, most likely contains images that depict the person who was queried in the Web search and thus represent the named entity in the text. In more detail, all features are compared with each other using a metric such as the cosine similarity. Let F_1 and F_2 be two different feature vectors, then the cosine similarity $s_{cos}(F_1, F_2)$ is defined as:

$$s_{cos}(F_1, F_2) = \left(\frac{F_1 \cdot F_2}{\|F_1\| \cdot \|F_2\|} \right) \tag{1}$$

The similarities among all feature comparisons are used to perform a hierarchical clustering that assigns images to the same cluster as long as their feature similarity is larger than a minimal similarity threshold τ_P. We calculate the mean of all feature vectors in the majority cluster to create a reference vector $\hat{F}_p$ for the person p. Finally, the feature vectors F_V of all faces $v \in V$ in the news image are compared to the reference vectors $\hat{F}_P$ of each person $p \in P$ mentioned in the text. Several options are available to calculate an overall *Cross-modal Person Similarity* (CMPS) such as the mean, n%-quantile, or the max of all comparisons. However, as mentioned above, usually the text contains more entities than the image and already a single correlation can theoretically ensure credibility. Thus, we define the *Cross-modal Person Similarity* (CMPS) as the maximum similarity among all comparisons according to Equation 2, since the mean or quantile would require

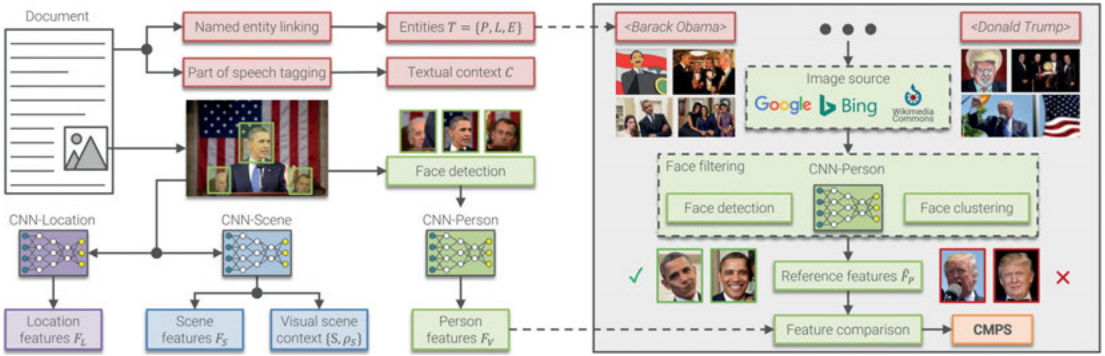


Fig. 8: Workflow of the proposed system. Left: Extraction of textual entities *T* and context *C* (red) as well as visual features *F* for persons (green), locations (purple), and scenes (blue). In addition, the visual scene context containing the probabilities ρ_S and word embeddings *S* of all scene classes is calculated. Right: Workflow to measure the *Cross-modal Person Similarity* between image and text. A similar pipeline is used for locations and events. All images (PD) by https://pixabay.com/service/license/.

the presence of several or all entities mentioned in the text.

$$\text{CMPS} = \max_{v \in V, p \in P} \left(s_{cos}(F_v, \hat{F}_p) \right) = \max_{v \in V, p \in P} \left(\frac{F_v \cdot \hat{F}_p}{\|F_v\| \cdot \|\hat{F}_p\|} \right) \tag{2}$$

This automatic system aims to support human assessors with measures of cross-modal entity consistency. In contrast to previous work, the system is completely unsupervised and does not rely on any predefined reference or training data. To the best of our knowledge, this is the first system *applicable to real-world news articles* by tackling several news-specific challenges such as the excessive length of news documents, entity diversity, and noisy reference (training) images downloaded from the Web that do not depict the target entity (e.g., images downloaded for *Barack Obama* might show additional or other persons). The applications are manifold, ranging from a retrieval system for news with low or high cross-modal coherence to an exploration tool that reveals the relations between image and text. The feasibility of our approach has been demonstrated on a novel large-scale dataset for cross-modal consistency verification that is derived from *BreakingNews* (Ramisa et al. 2018). The dataset contains real-world news articles in English and covers different topics and domains. The entities are manipulated with more sophisticated strategies than in previous work in order to obtain challenging datasets. The experiments show that our adaptive approach allows us to measure cross-modal consistency for arbitrary image-text pairs in news, without the need for any labeled training data.

Further details about the approach and experimental results are described in the corresponding paper (Müller-Budack et al. 2020). Source code and datasets are publicly available[2].

5 Conclusions

In this paper, we have presented computational approaches to model and automatically predict image-text relations and classes. For this purpose, we have introduced the dimensions (metrics) of *cross-modal mutual information*, *semantic correlation*, and the *status relation*. These dimensions allowed us to derive eight meaningful classes of image-text relations, which have been set in relation to existing taxonomies in the literature, particularly by Martinec & Salway (2005); McCloud (1993); Barthes (1977). We have also shown how to acquire training data for (deep) machine learning models. Additionally, we presented an overview of

[2] https://github.com/TIBHannover/cross-modal_entity_consistency.

a second approach, which allows us to measure different aspects of cross-modal mutual information for multimodal news in more detail, particularly with regard to cross-modal consistency. The aspects comprise (cross-modal) person similarity, location similarity, event similarity, and context similarity. Experimental results demonstrated the feasibility of both approaches.

Both approaches contribute to bridging the semantic gap. On the one hand, the definition of the metrics *cross-modal mutual information*, *semantic correlation*, and the *status relation* provides a differentiated computational model for the complex interplay of image and text. On the other hand, the second approach offers a differentiated perspective on different kinds of cross-modal mutual information, while suggesting metrics to measure cross-modal similarity for specific entity types. In future work, we plan to extend the latter approach also for the dimension of semantic correlation. Furthermore, the adaptation of our models to videos is an interesting direction of future research.

Bibliography

Aafaq, Nayyer, N. Akhtar, W. Liu, S. Z. Gilani & A. Mian. 2019. Spatio-Temporal Dynamics and Semantic Attribute Enriched Visual Encoding for Video Captioning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 12487–12496. Computer Vision Foundation/IEEE. 10.1109/CVPR.2019.01277. http://openaccess.thecvf.com/content_CVPR_2019/html/Aafaq_Spatio-Temporal_Dynamics_and_Semantic_Attribute_Enriched_Visual_Encoding_for_Video_CVPR_2019_paper.html (last accessed: 1 September 2021).

Abadi, Martín, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. J. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Józefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. G. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. A. Tucker, V. Vanhoucke, V. Vasudevan, F. B. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu & X. Zheng. 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *CoRR* abs/1603.04467. http://arxiv.org/abs/1603.04467 (last accessed: 1 September 2021).

Afouras, T., J. S. Chung, A. Senior, O. Vinyals & A. Zisserman. 2018. Deep Audio-visual Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1–1. https://doi.org/10.1109/TPAMI.2018.2889052.

Amodei, Dario, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, E. Elsen, J. H. Engel, L. Fan, C. Fougner, A. Y. Hannun, B. Jun, T. Han, P. LeGresley, X. Li, L. Lin, S. Narang, A. Y. Ng, S. Ozair, R. Prenger, S. Qian, J. Raiman, S. Satheesh, D. Seetapun, S. Sengupta, C. Wang, Y. Wang, Z. Wang, B. Xiao, Y. Xie, D. Yogatama, J. Zhan & Z. Zhu. 2016. Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin. In M. Balcan & K. Q. Weinberger (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New*

York City, NY, USA, June 19-24, 2016, vol. 48 JMLR Workshop and Conference Proceedings, 173–182. JMLR.org. http://proceedings.mlr.press/v48/amodei16.html (last accessed: 1 September 2021).

Anderson, Peter, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould & L. Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 6077–6086. IEEE Computer Society. https://doi.org/10.1109/CVPR.2018.00636.

Baltrusaitis, Tadas, C. Ahuja & L. Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41(2). 423–443. https://doi.org/10.1109/TPAMI.2018.2798607.

Barthes, Roland. 1977. *Image – Music – Text*. London: Fontana Press. Edited and translated by Stephen Heath.

Bateman, John. 2014. *Text and Image: A Critical Introduction to the Visual/Verbal Divide*. London: Routledge.

Bateman, John A., J. Wildfeuer & T. Hiippala. 2017. *Multimodality – Foundations, Research and Analysis. A Problem-Oriented Introduction*. Berlin: De Gruyter Mouton.

Benenson, Rodrigo, S. Popov & V. Ferrari. 2019. Large-Scale Interactive Object Segmentation With Human Annotators. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 11700–11709. Computer Vision Foundation/IEEE. https://doi.org/10.1109/CVPR.2019.01197.

Chorowski, Jan, D. Bahdanau, D. Serdyuk, K. Cho & Y. Bengio. 2015. Attention-Based Models for Speech Recognition. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama & R. Garnett (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 577–585. http://papers.nips.cc/paper/5847-attention-based-models-for-speech-recognition (last accessed: 1 September 2021).

Deng, Jia, W. Dong, R. Socher, L. Li, K. Li & F. Li. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, 248–255. IEEE Computer Society. https://doi.org/10.1109/CVPR.2009.5206848.

Deng, Jiankang, J. Guo, N. Xue & S. Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 4690–4699. Computer Vision Foundation/IEEE. https://doi.org/10.1109/CVPR.2019.00482.

Devlin, Jacob, M. Chang, K. Lee & K. Toutanova. 2019. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran & T. Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics. https://doi.org/10.18653/v1/n19-1423.

Ding, Changxing & D. Tao. 2018. Trunk-Branch Ensemble Convolutional Neural Networks for Video-Based Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 40(4). 1002–1014. https://doi.org/10.1109/TPAMI.2017.2700390.

Ewerth, Ralph, M. Springstein, L. A. Phan-Vogtmann & J. Schütze. 2017. “Are Machines Better Than Humans in Image Tagging?” - A User Study Adds to the Puzzle. In J. M. Jose, C. Hauff, I. S. Altingövde, D. Song, D. Albakour, S. N. K. Watt & J. Tait (eds.), *Advances in Information*

Retrieval - 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings, series = Lecture Notes in Computer Science, vol. 10193, 186–198. https://doi.org/10.1007/978-3-319-56608-5_15.

Halliday, Michael Alexander Kirkwood & C. M. Matthiessen. 2013. *Halliday's Introduction to Functional Grammar*. Oxfordshire: Routledge.

He, Kaiming, X. Zhang, S. Ren & J. Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778. IEEE Computer Society. https://doi.org/10.1109/CVPR.2016.90.

Henning, Christian Andreas & R. Ewerth. 2017. Estimating the Information Gap between Textual and Visual Representations. In B. Ionescu, N. Sebe, J. Feng, M. A. Larson, R. Lienhart & C. Snoek (eds.), *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ICMR 2017, Bucharest, Romania, June 6-9, 2017*, 14–22. ACM. https://doi.org/10.1145/3078971.3078991.

Henning, Christian Andreas & R. Ewerth. 2018. Estimating the Information Gap between Textual and Visual Representations. *Int. J. Multim. Inf. Retr.* 7(1). 43–56. https://doi.org/10.1007/s13735-017-0142-y.

Huang, Fei, Y. Cheng, C. Jin, Y. Zhang & T. Zhang. 2017. Deep Multimodal Embedding Model for Fine-grained Sketch-based Image Retrieval. In N. Kando, T. Sakai, H. Joho, H. Li, A. P. de Vries & R. W. White (eds.), *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, 929–932. ACM. https://doi.org/10.1145/3077136.3080681.

Huang, Ting-Hao (Kenneth), F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. B. Girshick, X. He, P. Kohli, D. Batra, C. L. Zitnick, D. Parikh, L. Vanderwende, M. Galley & M. Mitchell. 2016. Visual Storytelling. In K. Knight, A. Nenkova & O. Rambow (eds.), *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies San Diego California, USA, June 12-17, 2016*, 1233–1239. The Association for Computational Linguistics. https://doi.org/10.18653/v1/n16-1147.

Hussain, Zaeem, M. Zhang, X. Zhang, K. Ye, C. Thomas, Z. Agha, N. Ong & A. Kovashka. 2017. Automatic Understanding of Image and Video Advertisements. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 1100–1110. IEEE Computer Society. https://doi.org/10.1109/CVPR.2017.123.

Kahou, Samira Ebrahimi, X. Bouthillier, P. Lamblin, Ç. Gülçehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. N. Dauphin, N. Boulanger-Lewandowski, R. C. Ferrari, M. Mirza, D. Warde-Farley, A. C. Courville, P. Vincent, R. Memisevic, C. J. Pal & Y. Bengio. 2016. EmoNets: Multimodal Deep Learning Approaches for Emotion Recognition in Video. *J. Multimodal User Interfaces* 10(2). 99–111. https://doi.org/10.1007/s12193-015-0195-2.

Karpathy, Andrej & F. Li. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 3128–3137. IEEE Computer Society. https://doi.org/10.1109/CVPR.2015.7298932.

Kato, Kosuke, I. Ide, D. Deguchi & H. Murase. 2015. Generation of a Video Summary on a News Topic Based on SNS Responses to News Stories. In J. Redi & S. Rudinac (eds.), *Proceedings of the Fourth International Workshop on Crowdsourcing for Multimedia, CrowdMM 2015, Brisbane, Australia, October 30, 2015*, 21–26. ACM. https://doi.org/10.1145/2810188.2810189.

Krippendorff, Klaus. 1970. Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educational and Psychological Measurement* 30(1). 61–70.

Krizhevsky, Alex, I. Sutskever & G. E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou & K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, 1106–1114. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks (last accessed: 1 September 2021).

Kruk, Julia, J. Lubin, K. Sikka, X. Lin, D. Jurafsky & A. Divakaran. 2019. Integrating Text and Image: Determining Multimodal Document Intent in Instagram Posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 4621–4631. https://doi.org/10.18653/v1/D19-1469.

Lan, Zhen-zhong, L. Bao, S. Yu, W. Liu & A. G. Hauptmann. 2014. Multimedia Classification and Event Detection Using Double Fusion. *Multimedia Tools Appl.* 71(1). 333–347. https://doi.org/10.1007/s11042-013-1391-2.

van Leeuwen, Theo. 2005. *Introducing Social Semiotics*. London: Routledge.

Lin, Tsung-Yi, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár & C. L. Zitnick. 2014a. Microsoft COCO: Common Objects in Context. In D. J. Fleet, T. Pajdla, B. Schiele & T. Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, vol. 8693 Lecture Notes in Computer Science, 740–755. Springer. https://doi.org/10.1007/978-3-319-10602-1_48.

Lin, Tsung-Yi, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár & C. L. Zitnick. 2014b. Microsoft COCO: Common Objects in Context. In D. J. Fleet, T. Pajdla, B. Schiele & T. Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, vol. 8693 Lecture Notes in Computer Science, 740–755. Springer. https://doi.org/0.1007/978-3-319-10602-1_48.

Marsh, Emily E & M. D. White. 2003. A Taxonomy of Relationships between Images and Text. *Journal of Documentation* 59(6). 647–672.

Martinec, Radan & A. Salway. 2005. A System for Image-Text Relations in New (and Old) Media. *Visual Communication* 4(3). 337–371.

McCloud, Scott. 1993. *Understanding Comics: The Invisible Art*. New York: Harper Perennial.

Mikolov, Tomas, I. Sutskever, K. Chen, G. S. Corrado & J. Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, Z. Ghahramani & K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, 3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality (last accessed: 1 September 2021).

Müller-Budack, Eric, K. Pustu-Iren & R. Ewerth. 2018. Geoaddress Estimation of Photos Using a Hierarchical Model and Scene Classification. In V. Ferrari, M. Hebert, C. Sminchisescu & Y. Weiss (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII*, vol. 11216 Lecture Notes in Computer Science, 575–592. Springer. https://doi.org/10.1007/978-3-030-01258-8_35.

Müller-Budack, Eric, J. Theiner, S. Diering, M. Idahl & R. Ewerth. 2020. Multimodal Analytics for Real-world News using Measures of Cross-modal Entity Consistency. In C. Gurrin, B. Þ.

Jónsson, N. Kando, K. Schöffmann, Y. P. Chen & N. E. O'Connor (eds.), *Proceedings of the 2020 on International Conference on Multimedia Retrieval, ICMR 2020, Dublin, Ireland, June 8-11, 2020*, 16–25. ACM. https://doi.org/10.1145/3372278.3390670.

Ngiam, Jiquan, A. Khosla, M. Kim, J. Nam, H. Lee & A. Y. Ng. 2011. Multimodal Deep Learning. In L. Getoor & T. Scheffer (eds.), *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, 689–696. Omnipress.

Nöth, Winfried. 1995. *Handbook of Semiotics*. Bloomington: Indiana University Press.

Otto, Christian, S. Holzki & R. Ewerth. 2019a. "Is This an Example Image?" - Predicting the Relative Abstractness Level of Image and Text. In L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff & D. Hiemstra (eds.), *Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14-18, 2019, Proceedings, Part I*, vol. 11437 Lecture Notes in Computer Science, 711–725. Springer. https://doi.org/10.1007/978-3-030-15712-8_46.

Otto, Christian, M. Springstein, A. Anand & R. Ewerth. 2019b. Understanding, Categorizing and Predicting Semantic Image-Text Relations. In A. El-Saddik, A. D. Bimbo, Z. Zhang, A. G. Hauptmann, K. S. Candan, M. Bertini, L. Xie & X. Wei (eds.), *Proceedings of the 2019 on International Conference on Multimedia Retrieval, ICMR 2019, Ottawa, ON, Canada, June 10-13, 2019*, 168–176. ACM. https://doi.org/10.1145/3323873.3325049.

Otto, Christian, M. Springstein, A. Anand & R. Ewerth. 2020. Characterization and Classification of Semantic Image-Text Relations. *Int. J. Multim. Inf. Retr.* 9(1). 31–45. https://doi.org/10.1007/s13735-019-00187-6.

Ramisa, Arnau, F. Yan, F. Moreno-Noguer & K. Mikolajczyk. 2018. BreakingNews: Article Annotation by Image and Text Processing. *IEEE Trans. Pattern Anal. Mach. Intell.* 40(5). 1072–1085. https://doi.org/10.1109/TPAMI.2017.2721945.

Russakovsky, Olga, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg & F. Li. 2015. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* 115(3). 211–252. https://doi.org/10.1007/s11263-015-0816-y.

Schroff, Florian, D. Kalenichenko & J. Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 815–823. IEEE Computer Society. https://doi.org/10.1109/CVPR.2015.7298682.

Smeulders, Arnold W. M., M. Worring, S. Santini, A. Gupta & R. C. Jain. 2000. Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. Pattern Anal. Mach. Intell.* 22(12). 1349–1380. https://doi.org/10.1109/34.895972.

Taigman, Yaniv, M. Yang, M. Ranzato & L. Wolf. 2014. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, 1701–1708. IEEE Computer Society. https://doi.org/10.1109/CVPR.2014.220.

Unsworth, Len. 2007. Image/Text Relations and Intersemiosis: Towards Multimodal Text Description for Multiliteracies Education. In *Proceedings of the 33rd International Systemic Functional Congress*, 1165–1205. Sao Paolo, Brazil: Pontificia Universidade Catolica de Sao Paulo.

Vaswani, Ashish, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser & I. Polosukhin. 2017. Attention Is All You Need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan & R. Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Informa-*

tion Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need (last accessed: 1 September 2021).

Vinyals, Oriol, A. Toshev, S. Bengio & D. Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 3156–3164. IEEE Computer Society. https://doi.org/10.1109/CVPR.2015.7298935.

Weyand, Tobias, I. Kostrikov & J. Philbin. 2016. PlaNet – Photo Geoaddress with Convolutional Neural Networks. In B. Leibe, J. Matas, N. Sebe & M. Welling (eds.), *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII, series = Lecture Notes in Computer Science*, vol. 9912, 37–55. Springer. https://doi.org/10.1007/978-3-319-46484-8_3.

Xu, Kelvin, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel & Y. Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In F. R. Bach & D. M. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, vol. 37 JMLR Workshop and Conference Proceedings, 2048–2057. JMLR.org. http://proceedings.mlr.press/v37/xuc15.html (last accessed: 1 September 2021).

Ye, K., N. Honarvar Nazari, J. Hahn, Z. Hussain, M. Zhang & A. Kovashka. 2019. Interpreting the Rhetoric of Visual Advertisements. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1–1. https://doi.org/10.1109/TPAMI.2019.2947440.

Zhang, Ke, K. Grauman & F. Sha. 2018a. Retrospective Encoders for Video Summarization. In V. Ferrari, M. Hebert, C. Sminchisescu & Y. Weiss (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VIII*, vol. 11212 Lecture Notes in Computer Science, 391–408. Springer. https://doi.org/10.1007/978-3-030-01237-3_24.

Zhang, Mingda, R. Hwa & A. Kovashka. 2018b. Equal But Not The Same: Understanding the Implicit Relationship Between Persuasive Images and Text. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, 8. BMVA Press. http://bmvc2018.org/contents/papers/0228.pdf (last accessed: 01 September 2021).

Zhen, Liangli, P. Hu, X. Wang & D. Peng. 2019. Deep Supervised Cross-Modal Retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 10394–10403. Computer Vision Foundation/IEEE. https://doi.org/10.1109/CVPR.2019.01064.

Zhou, Bolei, À. Lapedriza, A. Khosla, A. Oliva & A. Torralba. 2018a. Places: A 10 Million Image Database for Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 40(6). 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009.

Zhou, Luowei, Y. Zhou, J. J. Corso, R. Socher & C. Xiong. 2018b. End-to-End Dense Video Captioning With Masked Transformer. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 8739–8748. IEEE Computer Society. https://doi.org/10.1109/CVPR.2018.00911.

Zoph, Barret, V. Vasudevan, J. Shlens & Q. V. Le. 2018. Learning Transferable Architectures for Scalable Image Recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 8697–8710. IEEE Computer Society. https://doi.org/10.1109/CVPR.2018.00907.

Part III: **Empirical Inroads: Case Studies and Results**

Andreas Rothenhöfer

“I can’t see why you’re laughing”: Multimodal Analysis of Emotionalized Political Debate

Squaring Multimodal Coherence, Body Reference and Facial Biometrics

Abstract: This case study focuses on various stages in which a sequence of interaction on TV is received on Twitter and subsequently co-constructed and reframed in a polarized way. In a mixed-methods approach, a biometric facial expression analysis has been applied in combination with established pragmatic approaches to reassess the issues at stake.

Keywords: language and emotion, social media, facial expression analysis, biometrics, multimodal interaction

1 Introduction: Novel Approaches to Analyzing Socially Co-Constructed Public Behavior

With the advancement of streaming, commenting, and sharing techniques that allow for an easy quotation of audio-visual content in microblogging platforms such as Twitter or Instagram, public attention to politicians’ embodied micro behavior has surged. Divisive and emotionalized TV debates, such as panel discussions on Brexit, are rich in expressive co-verbal and nonverbal interactions. This case study focuses on various stages in which a particular multimodal sequence of interaction on TV is socially received on Twitter and subsequently co-constructed and reframed in a polarized way. In order to help objectify the heart of the controversy and to explore novel approaches to analyzing, reconstructing, and evaluating the controversial sequence of events, a biometric facial expression analysis software (*iMOTIONS* biometric platform with *Affectiva* facial expressions analysis) has been applied in combination with established pragmatic approaches, so as to attempt a triangulation of qualitative pragmatic and quantitative biometric methods.

https://doi.org/10.1515/9783110725001-006

2 Wrong Words or Wrong Looks?

The media controversy was initiated by a remark made on the BBC's *Andrew Marr Show* on 13 October 2019[1]. In a remote interview setting, a life-size image of the UK's Home secretary Priti Patel is displayed on a monitor, while the host, facing the monitor, is sitting with his back to the camera so that the audience can observe the interviewee's immediate facial reactions to his questions. In the course of events, Patel became embroiled in a debate about displaying an allegedly inappropriate smile or smirk while Marr was reading out dire warnings about Brexit-related risks. Marr, on the other hand, became criticized for reprimanding Patel's 'natural' behavior in an inappropriate way, which later even caused the BBC to release an official apology on his behalf. Here is a verbal transcript of the interview passage in question, including lines by talk show host Andrew Marr (AM) and Priti Patel (PP):

AM (08:05)The government's own modelling suggests that this FTA would result in a lower growth of 6.7%. Is that something that you're prepared to accept as a price worth paying?
PP (08:16)Well, I don't accept that, and you know, I don't know which data you're quoting.
AM (08:20)It's your own government's figures.
PP (08:21)— and there's been a range — well, Andrew, there's been a range of information that has been put out in the media over recent weeks and months, much of which, I should say, is out of date. And …
AM (08:31)Okay, let's, lets …
PP (08:31)… as someone — that sits — someone who sits …
AM (08:32)… let's move away from the government then …
PP (08:32)… on the Cabinet committee every single day, where we're looking at the preparations for Britain post-Brexit, preparing for a no deal as well, you know, we have every confidence in our economy, in our businesses, but also in terms of future prospects post-Brexit.
AM (08:48)Ok, let's let's hear from those businesses directly then, because a whole bunch of them — and I'll read them out — the Society for Motor Manufacturers and Traders, the people that make and sell cars; the Chemical Industries

[1] The talkshow transcript is available at http://news.bbc.co.uk/2/shared/bsp/hi/pdfs/13101903.pdf, last accessed: 01 May 2020. The BBC's version has been rearranged in the above quotation block to better match the actual sequencing of overlapping turns in the video interview. The video is available in an unoffical recording at https://www.youtube.com/watch?v=ToRDRibXVOc, last accessed: 01 May 2020.

Association; the Food and Drink Federation; the Association of the British Pharmaceutical Industry, and the Aerospace Trade Industry body, a lot of people who are actually at the forefront of trying to make this country earn its place in the world, sent a letter to the government, which I will now read out part of to you. And they've said that this proposal: "is a serious risk to manufacturing competitiveness and will result in huge new costs and disruption to UK firms."

… (09:26)— **I can't see why you're laughing.** —

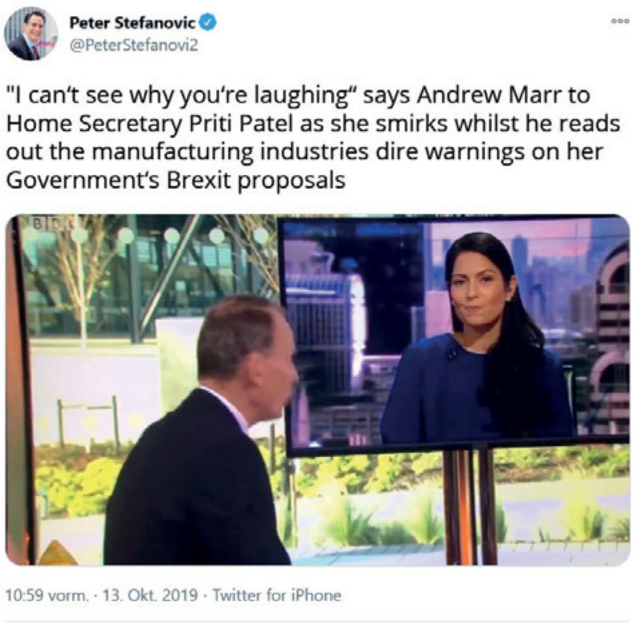
… (09:27)"It's got the potential to risk consumer and food safety and confidence, access to overseas markets for UK exporters and vital future investment in innovation in this country." That is a really serious challenge to this plan, is it not?

PP (09:40)Well, this is why the government has been working assiduously — with business as well I should say — across a range of sectors when it comes to planning for our exit from the European Union.

3 (Re-)Framing the Evidence on *Twitter*

Centre of the ensuing Twitter debate is the sequence of Patel's facial expressions preceding Marr's rebuke ("I can't see why you're laughing") and Marr's response to her behavior. Public attention is drawn to this ephemeral situation by Marr's explicit referencing and denoting Patel's facial expression as "laughing" in (9:26). Moreover, by using an authoritarian and evaluative construction *I can't see why you're X_V-ing* that indicates (or stages) moral indignation and seems typical for a hierarchical situation such as a head teacher telling off a pupil, Marr attempts to tighten the tone of the interview. This unusually emotionalized interaction and verbal highlighting of Patel's nonverbal behavior helps prepare the ground for turning the situation into a viral subject of political memefication and media discourse.

In the further sequence of events, image processing, social sharing, and commenting functions on Twitter allow the situation to be constantly reframed in line with either of two narratives along the Brexit divide. For one, journalist *@PeterStefanovi2* posts a video clip of the interview (see Figure 1; @PeterStefanovi2 2019). His slightly partial stance is subtly indicated by categorizing Patel's behavior as factually *smirking* (a category negatively connoted but situationally possibly less offensive than the more obtrusive and intentionally controllable notion of laughing) and by qualifying the industries' warnings as 'dire'. In Example (1), another user, *@Iancoll94354676*, then retweets *@PeterStefanovi2*'s original post as a quote, using "*smirking*" attributively, while reinforcing both the urgency of the economic situation and expressing his strong condemnation of Patel in an emotionalizing slur.



Peter Stefanovic
@PeterStefanovi2

"I can't see why you're laughing" says Andrew Marr to Home Secretary Priti Patel as she smirks whilst he reads out the manufacturing industries dire warnings on her Government's Brexit proposals

10:59 vorm. · 13. Okt. 2019 · Twitter for iPhone

Fig. 1: Original tweet by *@PeterStefanovi2* (2019).

> **(1)** *@Iancoll94354676* Marr to the smirking Patel about the very real threat to people's livelihoods, "**I can't see why you're laughing**" Patel is a vile excuse for a human being with no redeeming features (@Iancoll94354676 2019).

By creating a selective permanent manifestation of a dynamic and ephemeral situation, multimodal social media can shift the focus of attention to aspects of a situation that would otherwise not have been readily accessible for commenting or deemed worthy of public debate. Often these impressions can be classified in the tradition of Ekman as "micro expressions", as "facial expressions that occur within a fraction of a second" that are analysed as "involuntary emotional leakage that exposes a person's true emotion" (Ekman 2020).

On the side of Patel's critics and political opponents, the narrative of economic devastation, Marr's legitimate indignation, and Patel's allegedly deliberate or cold-hearted and habitual displays of haughtiness prevail (see Examples (2)–(4)). On the other side of the divide, the appropriateness of Marr's intervention, derogatorily classified as a rant, as well as *@PeterStefanovi2*'s selection or creation of evidence ("edited for maximum effect") is challenged (see Examples (5)–(6)). This stance is followed by further discussions on whether the category of rant is adequately ascribed to Marr (see Examples (8)–(9)).

(2) *@OilyBee* I only ever see her smirking. What a piece of work (@OilyBee 2019)

(3) *@JohnABrereton* I noticed you only showed Marr's rant and not Patel's reply. Edited for maximum effect, well done Peter, a true EU bigot (@JohnABrereton 2019a)

(4) *@FrankDoran9* I don't think it's smirking ' more likely to be rictus (@FrankDoran9 2019)

(5) *@WebMcnally* Where was the rant? Seriously, learn what words mean before you use them (@WebMcnally 2019)

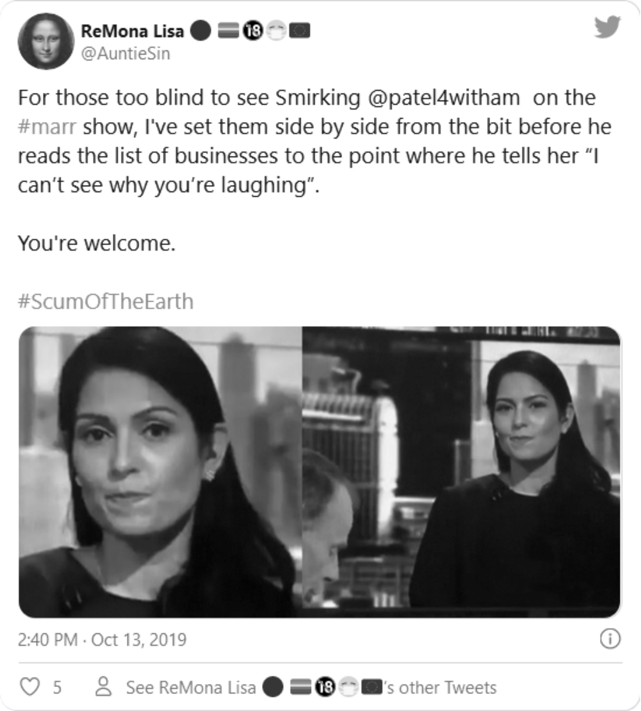
(6) *@JohnABrereton* Yes, you definitely should (@JohnABrereton 2019b)

(7) *@fski31* She always looks like she's just punched a cat and got away with it (@fski31 2019)

(8) *@DamCou* A BBC journalist saying "I can't see why you're laughing" to someone who isn't laughing is plain sinister. Even if she had been, I don't give a toss about how politicians *feel* about things — especially speculative "projections". I want to know what they're going to DO about them (@DamCou 2019)

(9) *@MrLukeGeorge* Andrew Marr tells Priti Patel "I can't see why you're laughing"//Except she wasn't! (@MrLukeGeorge 2019)

Users *@DamCou* and *@MrLukeGeorge* eventually reopen the evidence and select another still image of Patel, depicting a slightly different, more controlled and serious appearing micro-expression, with the corners of her mouth moved further down (see Examples (8)–(9)) while user *@AuntieSin* repeats *@DamCou*'s frame snapshot and combines it with yet another choice of frame snapshot, thereby contrasting two micro expressions with more and less raised corners of the mouth (see Figure 2). The user suggests that the "smirking" of Patel correlates with the expression "from the bit before he reads the list of businesses to the point where he tells her "I can't see why you're laughing", thereby indicating that it was only Marr's rebuke that eventually caused Patel to straighten her expression to the appearance captured by *@DamCou*.



ReMona Lisa ● ▬ ⑱ 😷 ■
@AuntieSin

For those too blind to see Smirking @patel4witham on the #marr show, I've set them side by side from the bit before he reads the list of businesses to the point where he tells her "I can't see why you're laughing".

You're welcome.

#ScumOfTheEarth

2:40 PM · Oct 13, 2019

♡ 5 See ReMona Lisa ● ▬ ⑱ 😷 ■'s other Tweets

Fig. 2: Tweet by *@AuntieSin* (2019) including contrasting snapshots.

4 Sequential Perception and Categorization: Hermeneutic vs. Biometric Approaches

Our previous discussion highlights a number of empirical and methodological challenges. Not only is the moral interpretation of events subject to divergent partial perceptions and therefore fiercely contested, but also the choice of visual evidence and the meaning of descriptive concepts. It is, however, striking, to what extent participants are ready to support their stance by providing visual evidence. Snapshots taken from dynamic sequences, e.g., a transitional still image of a person's volatile communicative, adaptor or facial expression behavior (cf. Ricci Bitti 2014), can be deliberately selected out of a variety of contradictory film frames so as to create a particular effect in line with the creator's own narrative. In our current context, one further challenge may be the exact conceptual meaning of *smirk* as a less prototypical and often evaluatively applied expression for a kind of smile. In many cases, attributions of untrustworthiness, affectedness or dishonesty are associated with the attribution of the verb. The Oxford English Dictionary does not provide any physiological features, but gives rather evaluative, mostly derogatory features and defines *smirk* as a subcategory to the generic term of *smile*:

> **smirk, v.** intransitive. To smile; in later use, to smile in an affected, self-satisfied, or silly manner; to simper. (OED Online 2020b)
> **smirk, n.** An affected or simpering smile; a silly, conceited, smiling look. (OED Online 2020a)

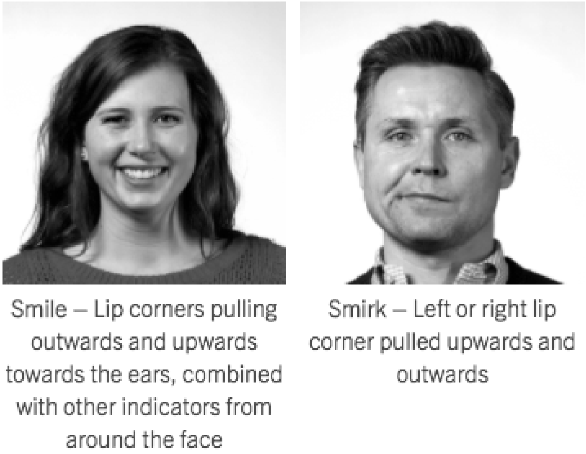
While there is little objectivity in morally evaluating human behavior, dynamic visual information can be captured and objectified with the help of automated detection tools. To what extent can biometric analysis tools replicate or support either of the two aforementioned narratives? Is there a correlation between behavior displayed by the interviewer and the interviewee? While Ekman's Facial Action Coding System (FACS) was a milestone in behavioral psychology as it provided a descriptive taxonomy of emotional expressions based on a system of universal facial action units, modern machine-learned biometric platforms can detect and process visual information very efficiently. However, the semantics of verbal descriptors may differ from what we find in general purpose dictionaries. Here, a smirk is technically defined predominantly in terms of its facial asymmetry:

> Asymmetric facial expressions, such as a smirk, are strong emotional signals indicating valence as well as discrete emotion states such as contempt, doubt, and defiance. Yet, the automated detection of asymmetric facial action units has been largely ignored to date. We present the first automated system for detecting spontaneous asymmetric lip movements as people watched online video commercials. Many of these expressions were subtle, fleeting, and co-occurred with head movements […]. (Senechal et al. 2013: 1)

Accordingly, in a blog by artificial intelligence software developer Affectiva dedicated to explaining the processes of emotion detection, mapping, and emotion metrics applied in their biometric software (Affectiva Blog 2017), the (prototypical) features of a smirk as opposed to a smile are illustrated in terms of their lip (a)symmetry (see Figure 3).[2]

Certain facial expression features are predictors for positive or negative emotion valence. Valence or hedonic tone is a descriptive dimension of an emotional state associated with a positively or negatively evaluated perception of an event, an object or situation (Frijda 1986: 207). It can therefore be an indication of a

[2] In similar terms, the current Unicode Emoji standard lists a symbol classified as Unicode Emoji No. 38 (U+1F60F) with an asymmetric, one-sided smile. It is comprised in the official Emoji table which assigns visual representations of Emojis with further descriptions.



Smile – Lip corners pulling outwards and upwards towards the ears, combined with other indicators from around the face

Smirk – Left or right lip corner pulled upwards and outwards

Fig. 3: Technical definitions of ‘smile’ and ‘smirk’ as presupposed in automated analysis (cf. Affectiva Blog 2017).

person’s perceived quality of interpersonal interaction, its topic, or the quality of a particular (temporary or habitual) relationship between interactants.

Tab. 1: Valence predictors implemented in Affectiva Mapping; (cf. Affectiva Blog 2017).

INCREASE POSITIVE LIKELIHOOD	INCREASE NEGATIVE LIKELIHOOD
smile	inner brow raise
cheek raise	brow furrow
	nose wrinkle
	upper lip raise
	lip corner depressor
	chin raise
	lip press
	lip suck

Tab. 2: Relationship between facial expressions and emotions predictors as implemented in Affectiva Blog (2017).

EMOTION	INCREASE LIKELIHOOD	DECREASE LIKELIHOOD
JOY	smile	brow raise brow furrow
ANGER	brow furrow lid tighten eye widen chin raise mouth open lip suck	inner brow raise brow raise smile
DISGUST	nose wrinkle upper lip raise	lip suck smile
CONTEMPT	brow furrow + smirk	smile

5 Multimodal Data Analysis and Interpretation

Let us now revisit the arguments of the contested Twitter-debate. The two contradictory perceptions of the interaction between Marr and Patel can be summarized as follows:

- *Position 1:* Marr inappropriately/condescendingly criticizes Patel's natural behavior with staged indignation in order to bring the politician out of her shell and to portray himself as a hard questioner.
- *Position 2:* Patel shows clear expressions of contempt and arrogance towards some very relevant, existential citizens' concerns raised by Marr and is therefore reprimanded for good reason.

While a biometric micro-analysis of the contested sequence of interaction cannot predict the audience's sympathies and whose performance they may more likely perceive as credible or plausible, it can help reconstruct and distinguish (causal) behavioral interaction chains in the dynamic foreground of events from the underlying general mood or attitude as a static background. With regard to the positions mentioned above, it may be crucial to determine, whether the biometric analysis will support the impression of a more positive smile (which may point towards cooperative friendly attitudes) or a smirk (that could be interpreted as an indication of non-cooperative and haughty contempt on Patel's part). Furthermore, anger or contempt may occur in a justifiable reaction to preceding face-threatening behavior by the interviewer, or it may be measured as some kind of underlying background attitude. The earlier case may be interpreted as unprofessional thinskinnedness, but not as a general negative or haughty attitude towards interviewer, interview or topic. Here Marr's rebuke may be blamed for an uneasy reaction on the interviewee's part. As Marr's face is not recorded by the camera, an interpretation of his potentially face-threatening behavior can only be inferred from the (para)verbal data of the interviewer's questions and remarks. A mixed methods approach comprises the following automated (1-3) and manual/hermeneutic (4-7) stages:

1. Automated detection of human faces (+ manual selection of a singular face box if necessary)
2. Quantitative Stage 1: biometric detection of facial landmark positions and relations within a face an their dynamics over a timeline (translated into facial expression measures)
3. Machine-learned combination of muscular action units into emotive gestalt bundles (translated into emotion measures) (all emotion assignments are probabilistic predications based on the co-occurrence of certain facial action features
4. Transcription/alignment of speech with facial action/emotion biometrics
5. Interpretation of interviewer's utterances as speech acts
6. Focus on situations of likely emotional significance (either in speech or image data)
7. Interpretation of speaker's/addressee's behavior as potential factors in an emotional interaction chain (emotional (dis-)agreement, face threatening behavior, etc.)

First (see Stages 1-3), a postproduction analysis of Patel's facial behavior was conducted for the contested monologic interview turn starting from the passage where Marr reads out a letter of business leaders and reprimands Patel for laughing. The analysis uses the iMOTIONS®Biometric Research Platform with an Affectiva®Facial Expression Analysis Engine.[3] The output diagram (see Figure 5) shows several biometric 'signal' tiers on a timeline. A first biometric analysis of the video-clip showed pronounced dynamics in a number of available signal tiers (derived by Affectiva from facial expression landmarks) that were therefore included: (1) dimpler, (2) lip press, (3) lip corner depressor and (4) lip pucker, (5) dimpler. Also the expressive bundles of (5) smile and (6) smirk and the emotion categories of (7) contempt, (8) disgust, (9) joy, and (10) anger have been included

[3] iMotions is an integrated platform solution for biometric behavioral research, used in various academic and commercial fields. It can combine different biometric sensors (EEG, GSR, eye tracking, automatic facial expression recognition).

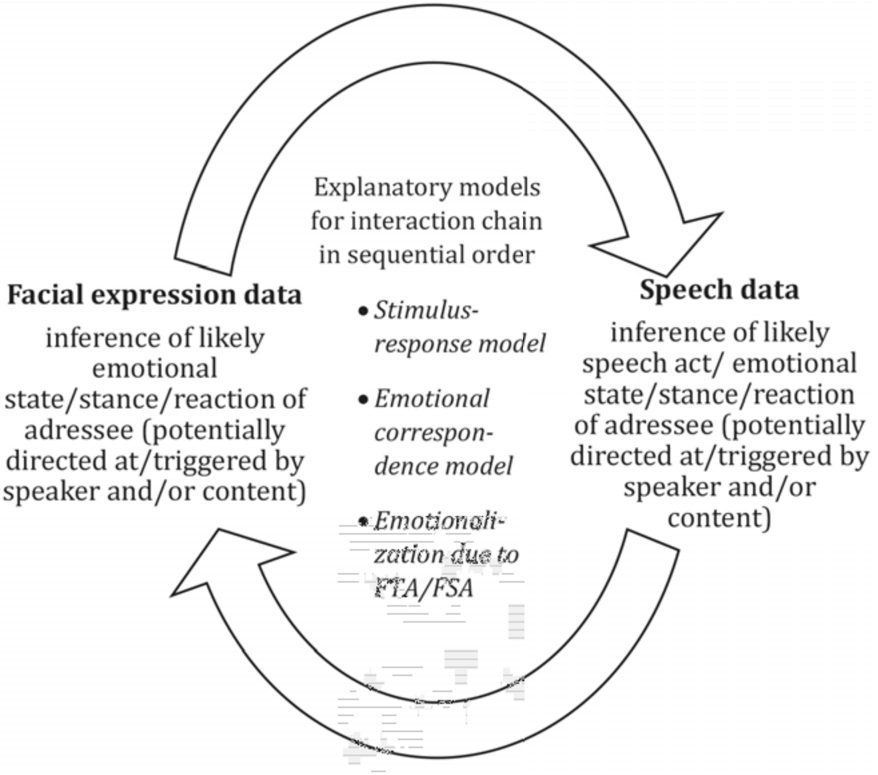
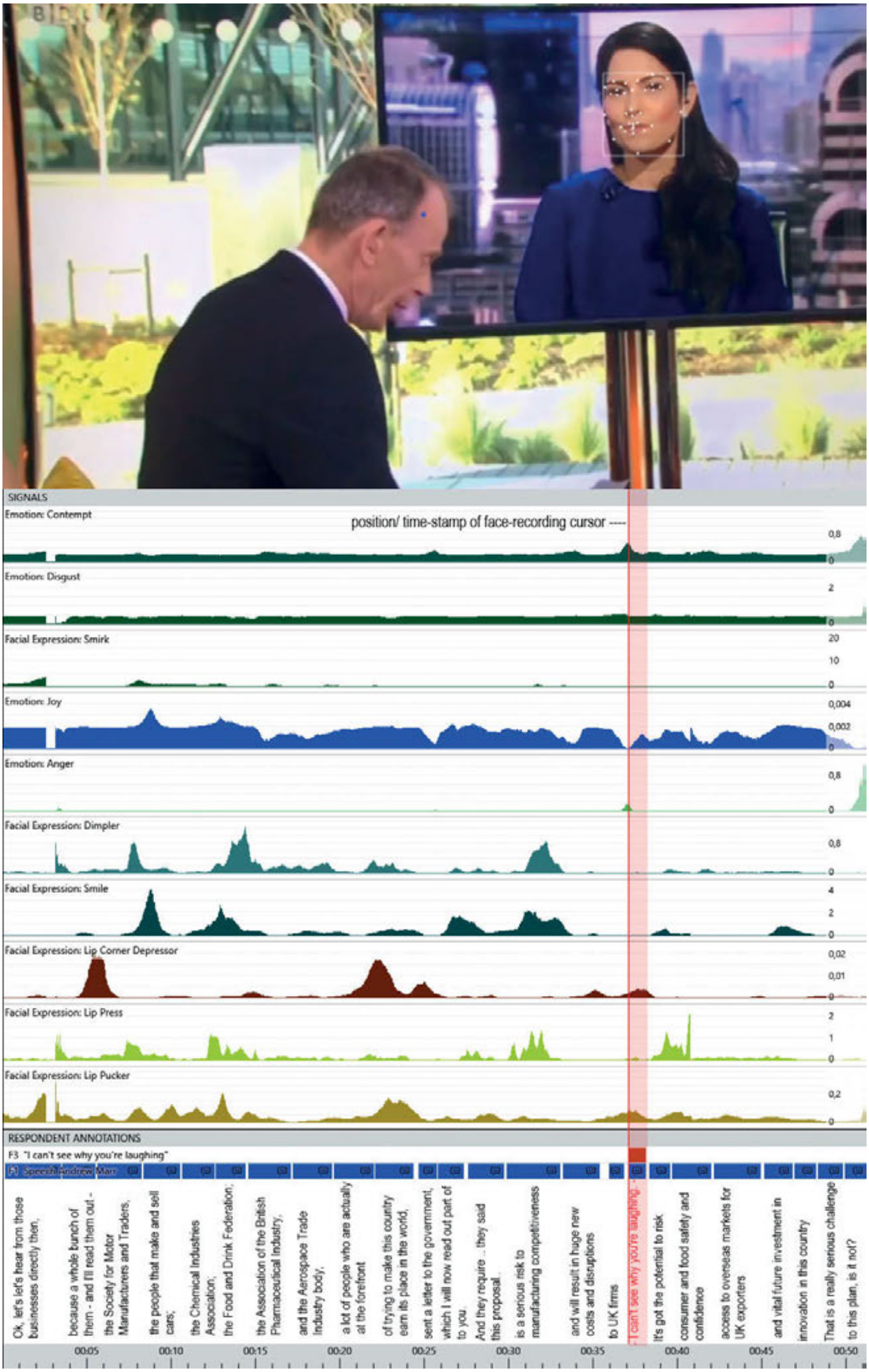

Explanatory models
for interaction chain
in sequential order
• Stimulus-response model
• Emotional correspondence model
• Emotionalization due to FTA/FSA
Facial expression data
inference of likely emotional state/stance/reaction of adressee (potentially directed at/triggered by speaker and/or content)
Speech data
inference of likely speech act/ emotional state/stance/reaction of adressee (potentially directed at/triggered by speaker and/or content)

Fig. 4: Hermeneutic circle of interpretation applied to the sequential correlation of facial expression data and speech data.

in the analysis. The reason for selecting the aforementioned signal features is that they either correlate with behavior (e.g., smile, smirk) or attitudes ascribed to Patel in social media discourse, e.g., haughtiness/ridicule vs. cooperative friendliness or with the likely emotional reaction to an ascribed face-threatening behavior by Marr (anger, contempt, doubt, defiance). Each signal output provided by Affectiva is autofocused in a bespoke scale related to the relevant movement range of a particular set of action units.

Just before the culmination in Marr's contested rebuke, there are several instances of stronger smiling displayed by Patel (27:560-35:433). These are accompanied by displays of disgust and contempt. Marr's strong rebuke ("I can't see why you're laughing!", 37:097-36:874) correlates with a delayed tightening of Patel's lip press action – which is most probably an intuitive impulse to better control her facial expression (and to avoid uncontrolled "derailment") and indicates a disrup-



F3 "I can't see why you're laughing"

F1 Speech Andrew Marr

Ok, let's let's hear from those businesses directly then,

because a whole bunch of them - and I'll read them out -

the Society for Motor Manufacturers and Traders,

the people that make and sell cars;

the Chemical Industries Association;

the Food and Drink Federation;

the Association of the British Pharmaceutical Industry,

and the Aerospace Trade Industry body,

a lot of people who are actually at the forefront

of trying to make this country earn its place in the world,

sent a letter to the government,

which I will now read out part of to you.

And they require ... they said this proposal

is a serious risk to manufacturing competitiveness

and will result in huge new costs and disruptions

to UK firms

- I can't see why you're laughing -

It's got the potential to risk

consumer and food safety and confidence

access to overseas markets for UK exporters

and vital future investment in

innovation in this country

That is a really serious challenge

to this plan, is it not?

Fig. 5: Face recording image (iMOTIONS 8.1 with Affectiva®Facial Expression Analysis Engine) with automated signal tiers and manual speech annotation.

tion of her perceived situational comfort (negative valence predictor, see Table 1). It correlates with an immediate decline of joy, a temporary increase in signaling of contempt and a moderate display of anger. Marr uses his face-threatening rebuke to alert Patel to his message and to provoke a more engaged reaction while he appears absorbed in the letter only a few seconds earlier. For the same reason, his intervention seems incoherent with his previous behavior, surprising, and by contrast rather harsh.

Our observations seem to prove both aforementioned social media positions partly right: There are indications of temporary derogatory smiling that were obviously registered by Marr and may be interpreted as the trigger for his intervention. However, such subtle signs of socially acceptable displays of smugness should be deliberately ignored by a professional interviewer. It would have been more advisable to lure the interlocutor out of her supposedly unempathetic façade with precisely targeted critical questions than to rise above her with face-threatening moral addresses that seem to backfire with parts of the audience and will thereby fail to hold the Home Secretary to account.

With regard to some of the initial ascriptions and assumptions in social media, the present analysis can deepen and objectify observations and show likely interdependencies between verbal and nonverbal interactions. Interviewer and interviewee certainly do not appear to find themselves in a casual relaxed environment. The situation is tense, not only due to policy challenges or complications and opposing public opinions. Also their professional roles and expectations (concerning party loyalty or journalistic impartiality) increase the complexity of interactions. Microexpressions as analyzed can provide interesting insights into subtle displays of inconsistency, authenticity, emotion, and temporary thin-skinned or unconscious losses of control in an overwhelmingly professionally controlled media setting. As a next step, including details on intonation and syllable stress, as well as the interviewer's own facial expressions, may contribute to an even more nuanced picture.

The mixed-methods approach relies on a correlation of quantitative data based on Patel's facial behavior with qualitative data (i.e., the manually correlated transcript of Marr's speech) as well as an annotation of speech act categories that may be interpreted as triggering an emotional reaction on Patel's side. Patel's expressions are interpreted as potential responses to the verbal and paraverbal stimuli originating in Marr's speech. Therefore verbal and facial information have to be organized along the same timeline.[4] The biometric video analysis has been

[4] iMotions' integrated export function allows for annotated images of the face-recording as well as the export of quantitative data in spreadsheets. Relevant human faces are automatically detected,

Tab. 3: Transcript aligned with percentual signal values and pragmatic annotation. As the range of muscular movement varies according to action unit, absolute metric measures provided by iMotions have been translated into relative percentages of prototypical action ranges. The autofocused maximum scale mark of each AU tier was taken to represent 100% of the prototypical range of motion, which explains peak values that may exceed 100%.

excerpt no	start time	end time	text	speech act/cooperation marker	Anger	Contempt	Disgust	Joy	Dimpler	Lip Corner Depressor	Lip Press	Lip Pucker	Smile	Smirk
1.	00:111	03:476	Ok, let's let's hear from those businesses directly then,	AUTHENTICITY, RELEVANCE	**1.9%**	26%	20%	5%	4%	1%	2%	11%	0%	3%
2.	03:500	05:500	because a whole bunch of them - and I'll read them out -	EMPHASIS, RELEVANCE	**1.7%**	25%	4%	4%	**31%**	1%	**22%**	11%	0%	0%
3.	05:524	08:269	the Society for Motor Manufacturers and Traders,	EMPHASIS, RELEVANCE	0.3%	24%	18%	5%	12%	**94%**	4%	10%	1%	0%
4.	08:293	10:547	the people that make and sell cars;	EMPHASIS, RELEVANCE	0.2%	26%	19%	6%	20%	0%	11%	**19%**	**40%**	**9%**
5.	10:571	12:571	the Chemical Industries Association;	EMPHASIS, RELEVANCE	0.3%	25%	21%	5%	2%	4%	1%	10%	2%	3%
6.	12:595	14:428	the Food and Drink Federation;	EMPHASIS, RELEVANCE	0.3%	24%	20%	6%	24%	1%	**51%**	**17%**	**31%**	4%
7.	14:547	17:086	the Association of the British Pharmaceutical Industry,	EMPHASIS, RELEVANCE	0.3%	24%	15%	5%	**102%**	11%	6%	**17%**	10%	1%
8.	17:166	19:570	and the Aerospace Trade Industry body,	EMPHASIS, RELEVANCE	0.4%	27%	19%	3%	16%	2%	4%	1%	2%	0%
9.	19:594	22:023	a lot of people who are actually at the forefront	RELEVANCE	0.3%	26%	20%	4%	13%	3%	2%	8%	7%	2%
10.	22:047	24:380	of trying to make this country earn its place in the world,	RELEVANCE, EMOTIONAL INVOLVEMENT	0.3%	25%	20%	5%	**38%**	**80%**	7%	13%	5%	2%
11.	24:632	25:743	sent a letter to the government,	CONTEXTUALISATION	0.3%	24%	21%	5%	2%	**31%**	0%	12%	8%	0%
12.	25:768	27:362	which I will now read out part of to you.	METACOMMUNICATIVE ANNOUNCEMENT, FOCUSSING ATTENTION	**1.5%**	**32%**	21%	2%	2%	8%	2%	6%	1%	0%
13.	27:560	29:719	And they require .. they said this proposal..	QUOTED EXPECTATION	0.2%	23%	19%	5%	4%	4%	**23%**	5%	**31%**	0%
14.	29:822	33:059	is a serious risk to manufacturing competitiveness	QUOTED WARNING	0.3%	24%	21%	4%	1%	1%	0%	3%	10%	0%
15.	33:187	35:433	and will result in huge new costs and disruptions	QUOTED WARNING	0.7%	29%	20%	2%	13%	2%	2%	2%	**26%**	2%
16.	35:930	36:874	to UK firms	EMPHASIS, RELEVANCE	0.3%	25%	22%	4%	0%	4%	0%	9%	1%	0%
17.	37:097	38:137	- I can't see why you're laughing. -	STRONG REBUKE	**17.2%**	**62%**	22%	0%	2%	10%	2%	**18%**	0%	0%
18.	38:288	39:613	It's got the potential to risk	MORE DIRECT WARNING	0.9%	**31%**	21%	2%	1%	12%	2%	6%	1%	0%
19.	39:637	42:031	consumer and food safety and confidence	RELEVANCE, AUDIENCE INVOLVEMENT	0.4%	26%	19%	4%	11%	1%	**32%**	13%	4%	0%
20.	42:111	44:983	access to overseas markets for UK exporters	RELEVANCE	1.1%	**34%**	19%	1%	6%	1%	3%	7%	1%	0%
21.	45:087	46:873	and vital future investment in	RELEVANCE	0.4%	27%	19%	3%	4%	2%	4%	8%	1%	0%
22.	46:898	48:246	innovation in this country	RELEVANCE	0.2%	23%	19%	5%	4%	1%	2%	4%	**14%**	0%
23.	48:271	49:855	That is a really serious challenge	WARNING, ADMONITION	0.3%	25%	20%	4%	2%	4%	1%	3%	1%	0%
24.	49:880	51:507	to this plan, is it not?	QUESTION TAG, DEMANDING FEEDBACK	1.0%	**32%**	18%	2%	2%	5%	2%	2%	0%	0%

manually annotated with Marr's questions in phrasal utterance blocks of approx. 2–3 min each.

Table 1 combines each sequence of the interviewer's utterance with the aforementioned quantitative signal data and aligns the initial time stamp of each utterance with the hearer's simultaneous facial signals.[5] The video has been analyzed processing every single video frame of 30fps. In order to ease qualitative comparison between significant parameters, all signals have been converted from absolute metric values to relative percentage values in line with the bespoke autofocus scales provided by the software for graph tiers (see Figure 5). Taking into account the linear quality of speech and the hearer's perceptive and cognitive processing lag, a potential facial reaction can be expected most likely towards the end of the segment or towards the beginning of the next segment. Furthermore, in the column 'speech act/cooperation marker' generic pragmatic categories of illocutionary forces or conversational maxims have been annotated that seem likely to cause or trigger a reaction or interaction from the hearer.

Between 03:500-2:2047[6] there are pronounced instances of dimpler and lip press action that may indicate characteristic reaction chains. While the dimpler seems less specific to valence prediction (see Table 1), lip press action is associated with an increased likelihood of negative valence (see Table 1). While the behavior bundle that is defined as smile is regularly measured before Marr's rebuke (08:293; 12:595; 27:560) and almost stops appearing afterwards (except for one instance around 46:898), a smirk (in line with the previously discussed technical definition of 'asymmetric smile') is not significantly displayed.[7] The strongest indications of smile can be observed following Marr's metapragmatic announcement of reading

traced, and marked with a white 'facebox' and facial landmark indicators which can be seen around and inside Patel's face in the recorded image. The red cursor line symbolizes the position of the face-recording at the time represented in the image, whereas the signals are displayed in tier graphs along the timeline. The verbal sequences aligned with the automatically detected facial signals have been annotated by hand, using an integrated annotation function. Its content cannot be displayed in the exported image; it was therefore manually mounted into the image and roughly aligned with the annotation blocks in vertical position (see Figure 5).

[5] Due to the organization of all available visual data along a processing frame rate of 30fps, the beginning of each utterance block has been aligned with the measured data of the closest matching video frame. As the illustrative spreadsheet in Table 1 does not cover all measured frames but specifically focusses on those related to utterance blocks, the measuring points cannot represent exact measures of extrema or points of inflection as displayed more adequately in the tier graphs in Figure 5.

[6] All numbers in the following section refer to the utterance starting time indicated in the first column of Table 1.

[7] It seems obvious that layman's everyday concepts of 'smirk' that were so dominant in Twitter discourse do diverge from the technical definitions applied above.

out the letter (08:293). Here, smiling may be explained as a cooperative or phatic signal of attentiveness or readiness and thereby be explained by quite the opposite of what Marr suggests with his rebuke. The dimpler as a more controlled indication of joy or amusement is particularly noticeable during the long and gravely read out litany of business and trade associations (03:500-02:204). As Patel's facial behavior does not mirror or show any resonance of Marr's gravitas, a certain cognitive or evaluative dissonance seems obvious. She may in fact not agree with Marr's aggravated tone or message. She may also be slightly absent-minded.

6 Conclusion

In political discourse, social emotionality and embodied perceptions are (co-)-constructed in an interactive, multimedial, and multimodal way and thereby fulfill central discourse functions. The multimodal structure of communication platforms such as Twitter can account for a conscious selection, manifestation, framing, and distribution of ephemeral aspects of "off the record" face-to-face communication, such as facial micro expressions, that are involved in a multi-stage selection, referencing, and interpretation process. Perceptions and interpretations of discourse visuals are turned into memes as relatable artifacts that are often constructed in a politically motivated, interest-driven way. The social meaning of verbal and nonverbal signs depends on a complex combination of semi-autonomous semiotic systems that should be analyzed from various vantage points.

Biometric analysis tools can support or contradict certain intuitions or perceptions in public discourse. Complex sequential correlations or causal dependencies in interactants' behavior can be traced and categorized and thereby contribute to a deeper and better informed multimodal interaction analysis. As things stand, especially in combination with qualitative approaches, the reliability of state-of-the-art tools like the one applied in the case study seem a valuable addition and data source for better informed and more holistic approaches to human interaction.

Our mixed-methods case study on Priti Patel's interaction with Andrew Marr was a novel attempt to combine selective emotionalized discourse adaptations and recontextualizations of media interaction with a close analysis of biometric and communicative data of the same event. Political framing of a person"s public interaction behavior in multimodal discourse representations can be complemented, challenged, and partly objectified by means of a systematic sequential facial expression and communication analysis. Causal interaction chains can only be determined by close analysis of verbal and nonverbal behavioral turns of all participants. Despite obvious remaining limitations of the existing tools and approaches, the method seems promising and should be further pursued and expanded.

Bibliography

Affectiva Blog. 2017. All About Emotion Detection and Affectiva's Emotion Metrics. https://blog.affectiva.com/emotion-ai-101-all-about-emotion-detection-and-affectivas-emotion-metrics (last accessed: 16 December 2020).

@DamCou. 2019. A BBC journalist saying "I can't see why you're laughing" to someone who isn't laughing is plain sinister. Even if she had been, I don't give a toss about how politicians *feel* about things—especially speculative "projections". I want to know what they're going to DO about them. 13 October 2019, 7:15 pm. https://twitter.com/DamCou/status/1183431190130429953.

Ekman, Paul. 2020. Micro Expressions. https://www.paulekman.com/resources/micro-expressions (last accessed: 16 December 2020).

@FrankDoran9. 2019. I don't think it's 'smirking' more likely to be rictus :-). 13 October 2019, 12:19 pm. https://twitter.com/FrankDoran9/status/1183326303614885888.

Frijda, Nico H. 1986. *The Emotions*. Cambridge: Cambridge University Press.

@fski31. 2019. She always looks like she's just punched a cat and got away with it. 13 October 2019, 12:27 pm. https://twitter.com/fski31/status/1183328400926220288.

@Iancoll94354676. 2019. Marr to the smirking Patel about the very real threat to people's livelihoods, "I can't see why you're laughing" Patel is a vile excuse for a human being with no redeeming features. 13 October 2019, 11:39 pm. https://twitter.com/Iancoll94354676/status/1183367876411101185.

@JohnABrereton. 2019a. I noticed you only showed Marrs rant and not Patels reply. Edited for maximum effect, well done Peter, a true EU bigot. 13 October 2019, 11:39 AM. https://twitter.com/JohnABrereton/status/1183316250191118336.

@JohnABrereton. 2019b. Yes, you definitely should! 13 October 2019, 06:44 AM. https://twitter.com/JohnABrereton/status/1183423372337864705.

@MrLukeGeorge. 2019. Andrew Marr tells Priti Patel "I can't see why you're laughing"//Except she wasn't! 13 October 2019, account/tweet not retrievable.

OED Online. 2020a. smirk, n. http://www.oed.com/view/Entry/182619 (last accessed: 16 December 2020).

OED Online. 2020b. smirk, v. http://www.oed.com/view/Entry/182621 (last accessed: 16 December 2020).

@OilyBee. 2019. I only ever see her smirking. What a piece of work. 13 October 2019, 11:15 AM. https://twitter.com/OilyBee/status/1183310355348496386.

@PeterStefanovi2. 2019. "I can't see why you're laughing" says Andrew Marr to Home Secretary Priti Patel as she smirks whilst he reads out the manufacturing industries dire warnings on her Government's Brexit proposals. 13 October 2019, 10:59 AM. https://twitter.com/PeterStefanovi2/status/1183306306372952064.

Ricci Bitti, Pio E. 2014. Facial Expression and Social Interaction. In C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill & S. Teßendorf (eds.), *Body – Language – Communication.*

(Handbooks of Linguistics and Communication Science (HSK), 38,1), 1342–1349. Berlin: De Gruyter Mouton.

Senechal, Thibaud, J. Turcot & R. el Kaliouby. 2013. Smile or Smirk? Automatic Detection of Spontaneous Asymmetric Smiles to Understand Viewer Experience. In *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 1–8. Shanghai: IEEE. https://www.affectiva.com/wp-content/uploads/2017/03/Smile_or_Smirk_Automatic_Detection_of_Spontaneous_Asymmetric_Smiles_to_Understand_Vi.pdf (last accessed: 1 May 2020).

@WebMcnally. 2019. Where was the rant? Seriously, learn what words mean before you use them. 13 October 2019, 02:42 PM. https://twitter.com/WebMcnally/status/1183362478039.

Christian Mosbæk Johannessen, Mads Lomholt Tvede, Kristoffer Claussen Boesen, and Tuomo Hiippala

A Corpus-Based Approach to Color, Shape, and Typography in Logos

Abstract: In this chapter our aim is to develop a corpus-based procedure for empirical studies of style in corporate logos. We present a data-driven corpus study of n=50 logos, 25 each from the oil industry and non-governmental environmental organizations (NGOs), in order to see if we can pick up significant differences in style between the two groups. We use a quantitative operationalization of distinctive features of color, shape, and typography, and measure how these are distributed in the logos. We then use principal component analysis to understand how these distinctive feature variables interact in each group. Our results show a subtle but statistically robust difference between oil company and NGO logos. For example, oil company icotypes tend to prefer straight and angular shapes and more saturated colors, NGO icotypes prefer curves and more subdued colors. Although our study does not warrant too much speculation about the values and ideologies underpinning these choices of visual style in the logos, we do find them sufficiently robust to justify a qualitative follow-up study seeking to connect our findings with the common notion that "brands are people" and so with social values of, for example, nature versus culture, with aesthetics of softness versus hardness, ideologies of altruism versus commercialism, to name a few.

Keywords: logo, corpus, graphic design, typography, principal component analysis

1 Introduction

This chapter presents an exploratory, data-driven corpus study of color, shape, and typography in the kind of thing we commonly refer to as *a logo*, the simple "multimodal ensemble" (following Jewitt 2013: 254–255) of verbal, typographic, and pictorial elements used by a company, organization or individual for purposes of identification and branding. A systematic review of everything that has been written about logos is beyond the scope of the chapter and difficult to achieve because many professional practices including graphic design (e.g., Adams et al. 2006; Airey 2015; Mollerup 1997), marketing (e.g., Bishop 2001; Llorente-Barroso & Garciaa-Garcia 2015), branding (e.g., Heilbrunn 1997, 2001), and intellectual property law (e.g., Beebe 2004), all share an interest in logos and each have developed their own terminology: marks, trademarks, signatures, word marks, device marks.

https://doi.org/10.1515/9783110725001-007

Semiotic treatments of logos as artefacts tends to be conceptual (Johannessen 2011; Skaggs 2017, 2018), discussing empirical examples mostly as illustrations of conceptual points, or qualitative single case studies (Cowin & Matusitz 2011; Johannessen 2016; Aiello 2017), or comparisons of just a few logos (Scott 1993; Johannessen 2017).

We think that logos ought to receive more attention in multimodality studies. They are a good example of a structurally simple multimodal ensemble the study of which would be very informative, but which is nevertheless often taken as a given in multimodal studies of higher-order phenomena such as text-image relations, multimodal discourse structure, etc. of which they are often a part. And because logos are, at least on casual inspection, quite simple, we can hope to achieve good descriptive exhaustiveness of very low-level structural features, and so learn something about multimodal artifacts beyond the study of logos. Yet, to our knowledge, no corpus-based, quantitative studies of logos exist.

2 Theoretical Background

2.1 Style as a Dimension of Semiotic Analysis

It is commonly accepted among logo designers that the best logos are simple and striking, an idea that has been emphasized again and again by several of the most influential architects, designers, and graphic designers in history. Quotes like Mies van der Rohe's "Less is More" and Louis Danziger's "If you can't explain the idea in one sentence over the telephone, it won't work" abound in graphic design books and blogs. Whether visual parsimony in fact makes a logo better remains an open question, but it seems to be a matter of accepted wisdom among graphic designers that "[...] simplicity also makes your design easier to recognize, so it stands a greater chance of achieving a timeless, enduring quality" (Airey 2015: 23).

The result of this preference for simplicity in both concept and execution is that logos have a very narrow informational "bandwidth" understood as differences that can make differences. They cannot denote[1] very elaborate or detailed concepts. Instead, they are apt for conveying what designers call the "look and feel" of a host's

[1] Hjemlslev's classical semiotic concepts of denotation and connotation are useful here. Denotation has been glossed by logo semiotician Steven Skaggs (2017: 256) as "The direct—often coded—convergent referent in the semantic register". He glosses its counterpart, connotation, as "The associative and allusive referents in the semantic register".

brand, the connotative potential that makes a logo “immediately recognizable [and] expresses a point of view” (Wheeler 2013: 148).

In this chapter, we think of look and feel in terms of visual style. Bateman et al. (2019: 8) assert that the empirical support for concepts of visual style is underdeveloped, and our work should be seen as part of an effort to understand it better. Our study is conceptually framed by Theo van Leeuwen’s (2005a) discussion of style in *Introducing Social Semiotics*, which he begins with this gloss of the term from the Concise Oxford Dictionary: “[Style is] a manner of writing, speaking, or doing, especially as contrasted with the matter to be expressed or thing done”. In these terms, we emphasize ‘manner of expression’ by looking only at formal variables of shape, color, and typography, and de-emphasize everything to do with the ‘matter expressed’ by those forms.

Van Leeuwen approaches style from three different perspectives, all of which have an individual-centric scope: (1) The idiosyncratic mannerisms that belong to the individual, and which can be both consciously performative or entirely outside conscious control, (2) the social style that marks social allegiance or solidarity, and (3) lifestyle, which is poised somewhere between individual and social styles and is, above all, an expression of consumer identity. We would add that style, insofar as style expresses identity, is not only an individual endeavor. Guided by the metaphor that “brands are people” (Koller 2009: 45), organizations express identities too. And not only does an organization express its individual identity much in the same way as a person can express “lifestyle” through a “composite of connotations” (van Leeuwen 2005a: 146), organizations can express something akin to “social style”, an allegiance to an industry or business sector, as well. It is this latter phenomenon which guides our study. Insofar as logos are more apt for specifying a connotational, stylistic look and feel than they are for denotative content, and insofar as organizations cluster according to the kind of business sector they are in, our goal is to find out whether we can operationalize the analysis of logos, in terms of their shape, color, and typography, in a way that could empirically address otherwise impressionistic claims that logos from different industries have different looks and feels.

2.2 Distinctive Features of Graphics

The term *graphics* is commonly used to refer to a very broad range of visual communication phenomena ranging from computer games and computer-generated images to lino cuts. We restrict our use of the term to denote a family of hapto-visual materialities, or “canvases” (Bateman 2008: 16), that share the affordance of dividing a surface into binary figure-ground regions with very distinct demarcations.

This basic feature of graphics harks back to the material affordances of early relief and intaglio print-making (in both of which material is carved, cut, corroded, etc. from a print matrix to yield areas that are either intact or not). Even today such a binary structuring of a two-dimensional space into figure and ground is a key principle in vector graphics applications such as Adobe Illustrator and FontLab, which use vector paths to demarcate the envelope of a figure from the ground. This family of canvases is used in expressive forms including, among others, writing, typography, cartography, the line art used in diagrams, pictograms, signage, app icons, to name a few. An effort to outline their formal commonalities has been made by Andreas Stötzner (2003) under the headline of "Signography".

Because logos are designed with simplicity of both form and content in mind, we suggest that our operationalization must begin with the most basic, formal levels of distinctions in the graphic canvases. This is important in order to avoid any circularity in our reasoning. We are attempting an operationalization of material properties of graphic canvases without any prior assumptions about how they might be perceived or make meaning (whether denotative or connotative). Only then can we begin to see how one industry may deploy color, shape, and typography in order to set itself visually apart from other industries.

There have been several attempts to describe graphic form in a way that makes it amenable to analysis, most notably in French cartographer and information designer, Jacques Bertin's (1983) *Semiology of Graphics*. Bertin (1983: 42–43) boils what he calls "the graphic system" down to eight variables a designer has to work with. In addition to the two planar dimensions, they are: (1) size, (2) value (similar to "brightness" in Photoshop's HSB-model of color, which describes a given color by its 'hue', 'saturation', and 'brightness'), (3) texture, (4) color, (5) orientation, and (6) shape. A similar intuition seems to guide some of the work undertaken by Gunther Kress and Theo van Leeuwen's "distinctive feature" approach to color (2002) and Johannessen's approach to graphic shape (2016). Because it has been so emphatically argued that typography is a rich resource for brand identity expression (Hyndman 2016), we have included elements from van Leeuwen's distinctive feature approach to typography (2005b; 2006) in our framework as well.

2.2.1 Distinctive Features of Color

In their article exploring color as a semiotic mode in its own right, Kress and van Leeuwen distinguish two types of affordance in color, "two sources for making meaning with color" (Kress & van Leeuwen 2002: 355). One is the association, or, in keeping with our own use of these concepts, the connotations of "look and feel". The other is the color's *distinctive features* (Kress & van Leeuwen 2002: 355) (inspired by Jakobson and Halle's distinctive feature phonology). They describe these as a range of formal parameters including the color's (1) hue (its position on a color wheel going through the visible spectrum from red to violet), (2) saturation (a scale from maximally saturated to gray), (3) value (a scale of brightness from white to black), (4) modulation (a scale from flat, unmodulated color to highly modulated showing many shades and tints of the same hue), (5) differentiation (a scale from monochromatic to multicolored with many hues), (6) purity (a scale from "pure" primary colors to "hybrid" secondary or tertiary colors).

2.2.2 Distinctive Features of Shape

Drawing inspiration from Kress and van Leeuwen's work on distinctive features, Johannessen (2016) has made an attempt at operationalizing shape characteristics of graphic canvases. He enumerates (1) straightness (the choice between straight and un-straight), (2) bend (the choice between angles and curves) and (3) direction (the choice between concave and convex), but stresses that a fractal-derived approach to these features of shape (that curves can be embedded in curves) is necessary in order to fully describe shape.

2.2.3 Distinctive Features of Typography

Finally, following his work with Gunther Kress on distinctive features of color, Theo van Leeuwen (2005b) has added his distinctive feature perspective to an already considerable body of literature on typography (see e.g., Gill 1988; Reimar & Birkvig 2003; Baines & Haslam 2005; Lupton 2014), attempting to boil typography down to a range of parameters on which, he argues, we can begin to discuss how typography can make meaning. These parameters are: (1) weight (the relative width of the strokes in letters), (2) expansion (the width of letterforms relative to their height), (3) slope (the angle of letters relative to their baseline), (4) curvature (the overall shape of letterforms in terms of angularity or smoothness), (5) connectivity (the connectedness of strokes in and between letterforms), (6) orientation (the height of letterforms relative to their width), and (7) regularity (the evenness of distribution of any structural feature).

3 Framework and Analytical Procedure

3.1 Logotypes and Icotypes

At its most general level of description, inspired by marketing semiotician Benoit Heilbrunn (1997: 177), our framework distinguishes two different kinds of elements, logotypes and icotypes, that can be combined into a logo. A *logotype* consists of symbols from writing systems, including numerals, and expresses what in trademark law is commonly referred to as a "word mark"; the proper name of the legal entity which is the logo's host. An *icotype* can consist of any graphic sign, but is usually a stylized depiction of some sort, and is referred to in trademark law as a "device mark". The far left of Figure 1 illustrates how a fictional logo for a host called "Ogol" can be broken into an icotype and a logotype.

3.2 Shape Analysis

The first step in our analysis of the shape of each logo is to determine the number of distinct bounded regions (or areas) in its logotype and icotype (insofar as they are both present). This step is illustrated in the middle of Figure 1. Our approach to shape does not regard the building blocks of letters in a logotype as lines or strokes (as in the analysis of typography described in Section 3.4) but as two-dimensional regions. As a consequence, the so-called "counters", the negative shapes inside the loops of letters such as 'o' and 'g' in the Ogol logo are also considered to be individual regions as illustrated in Figure 1. In the example of the Ogol logo, there are 7 regions in the icotype and 5 in the logotype. To be clear, the analysis operationalizes certain material properties of space and shape in graphic canvases at a purely formal level. It does not model how we in fact perceive a logo (for example, we perceive the Ogol icotype as three overlapping, transparent shapes fanned out around a hole, not 7 individual graphic regions), nor how it could be said to be meaningful.

Our analysis of shape only looks at instances of 'straight', 'angle', and 'curve', as we assumed these to be more telling of visual style than whether a shape feature was 'convex' or 'concave'. Thus, the second step is to count all instances of three types of shape — straights, angles, and curves — for all regions in the icotype and logotype. The approach is inspired by how shapes are described in vector graphics and is illustrated on the right in Figure 1. We count an occurrence of straight, angle or curve where we would expect them in a Bezier-based vector description of that shape. The actual annotation is done by importing a bitmap image of the

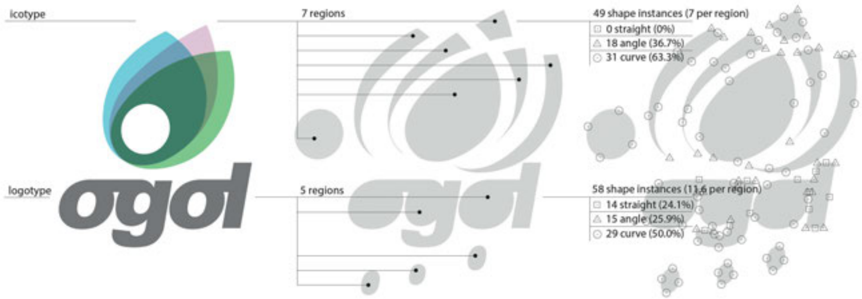


Fig. 1: Illustration of the principle behind breaking an icotype and logotype into distinct graphic regions and analyzing these in terms of shape instances.

logo into Adobe Photoshop and using the "count tool" to keep track of shape feature occurrences (see Figure 5). It should be noted that, whereas annotating occurrences of straight and angle is fairly straightforward, curves are harder to distinguish, especially in cases where one curve transitions smoothly into another[2] In the fictional Ogol icotype there are 49 instances of shape-defining nodes, 7 on average per region. 0% of these are straight, 36.7% are angles and the majority, 63.3%, are curves. In the logotype there are 58 instances of shape features, 11.6 on average per region, 24.1% straight, 25.9% angles, and 50% curves.

This approach enables us to quantify how shape features are distributed in the logo. A key metric is the shape density; basically a count of instances of shape features. The Ogol icotype has a density of 49, the logotype 58. In itself, this says very little. However, if we average densities over regions, we arrive at a potentially much more interesting number, the density per region (which we will refer to as 'D/r' in what follows): $\frac{sum\ of\ densities}{regions}$. This characterizes the overall distribution of shape features in the logo far more usefully. The Ogol icotype has a D/r of 7.0, the logotype 11.6.

D/r is potentially interesting because, as illustrated by Figure 2, the visual style of a logo can come across as clean or grainy depending on how many instances of shape features each region contains.

[2] Ideally, we would have collected vector files of the 50 logos in order to draw on the vector nodes used in their rendering to make such distinctions. In practice it has not been possible to obtain vector files of all the logos, and we have chosen to work from available bitmap images instead and accept that, in coding adjacent curves, we had to rely on our judgment to tell where one curve ended and the next began.

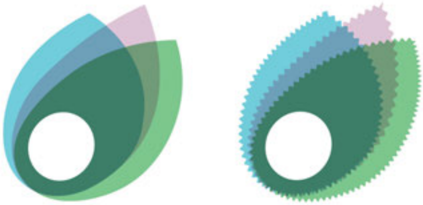


Fig. 2: Illustration of the different look and feel of the Ogol icotype in the original (left) and with a zig-zag filter, 30 ridges per line segment, applied in Illustrator (right) to increase the density per region (D/r). Note that, at one level, the regions retain their leafy shapes. But on a lower, embedded level, they take on a grainier look and feel. The material ability of graphic canvases to embed shape within shape (within shape) suggests a fractal, self-similar organization of shape characteristics.

Our approach also enables us to quantify the proportion of straights, angles, and curves to the total number of shape features, $\frac{shape\ type}{density}$, of an icotype or logotype (we refer to these metrics as S/D, A/D, and C/D). In the Ogol logo, for example, the majority of shape features are of the curve type (63.3% in the icotype, 50% in the logotype). We express these proportions as values between 0 and 1.

3.3 Color Analysis

As we stated in the introduction, this study is exploratory and data-driven. In its course, we have decided to exclude some of the distinctive features presented in Section 2. Our analysis of color only looked at 'hue', 'saturation', and 'value' (and not 'modulation', 'differentiation', and 'purity') because these are conveniently operationalized using Photoshop's "eyedropper tool" and the "HSB" model of color (which stands for hue, saturation, brightness). The software expresses hue on a range from 0–360° (as a position on a color wheel with bright red at 0°, or 12 o'clock if it were a clock dial). Saturation and brightness, however, are expressed by Photoshop on a scale from 0% to 100%.

Our analysis of color is deliberately kept very simple. We import a bitmap image of each logo into Adobe Photoshop and use the eyedropper tool to measure the HSB values for each region in the icotype and logotype as illustrated by Figure 3. This enables us to average the values for color saturation and color brightness over each icotype and logotype. The Ogol icotype is of fairly low color saturation: only 27.7% on average. This translates to colors that are bled out, grey-ish and subdued as opposed to vibrant. It is, however, also fairly bright: 74% on average (as opposed to the darker logotype, which is only 47%). Note that regions, which we take to be transparent, such as the loops in the letters ‘o’ and ‘g’ in Ogol, are not included in the analysis.

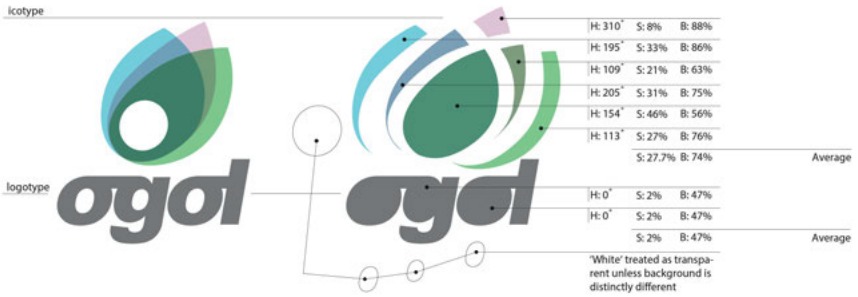


Fig. 3: Illustration of the principle used to arrive at the average color saturation and brightness of each icotype and logotype. For each distinctly colored region a measure of hue, saturation, and brightness is taken. We then average over saturation and brightness for icotypes and logotypes respectively. Note that enveloped regions that are ‘white’ are treated as negative shapes (when the background is also white), and so transparent, or colorless. Examples are the ‘hole’ in the icotype’s leaf shapes and the ‘holes’ in the logotype’s letters.

Because there is no meaningful way to take an average over the hue values, we have excluded this measurement from our study. In future studies, however, we may choose to group observed hue values in, for example, 45° intervals, in order to compare the proportions of colors in different intervals across industries. Furthermore, the two samples show only little variation in hue but significant variation in saturation and brightness, which makes these features more interesting for our comparison.

3.4 Typographical Analysis

Icotypes are predominantly pictorial and logotypes are predominantly verbal and typographic. Even if one sometimes bleeds into the other, they are different canvases. The same material may serve as a base for both and their shape and color can be analyzed according to the same procedure, but they are not identical in all respects. As a result, the fundamental units of our operationalization of typography of logotypes are those common in the type industry: letters and strokes (recall that in Section 3.2 on ‘shape’, we regarded them as regions).

Although we cannot rule out the possibility of type occurring as an integrated part of the icotype in some logos, we choose only to apply the typographic analysis on the logotype. In operationalizing van Leeuwen's distinctive features of typography, we have taken a few liberties. The 'curvature' of letterforms is captured in our analysis of shape, so it is not treated here. Also, we have fused van Leeuwen's 'expansion' and 'orientation' under the label 'expansion'. We found the two features to be redundant: expansion can be described as the width of letterforms relative to their height, and orientation is the reverse, that is, height relative to width. Finally, we disregarded the feature 'regularity'. Regularity can be operationalized with respect to any feature of visual structure, not only of typography, and we simply have not come up with a clever way of capturing it. Maybe, in future studies, one could apply measures of entropy in pixel distributions of bitmap images (e.g., Silva et al. 2016).

The first step in our typographical analysis is determining the number of letters (and also spaces between letters) as well as the number of strokes those letters are made up of. The second step is to import a bitmap of the logo into Adobe Illustrator and use the "measure tool" to measure lengths and angles as illustrated in Figure 4.

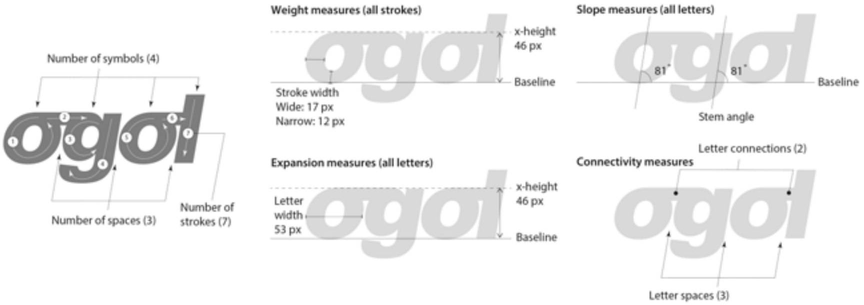


Fig. 4: Illustration of the measures taken of each logotype in order to analyze distinctive features of typography.

3.4.1 Weight

In order to measure the weight of letters, we have adapted Johannessen's (2011: 244) "weight scale rating" (WSR) to express weight as a value between 0 and 1 by relating stroke width to stroke length (and because stroke lengths vary, we use x-height as a proxy[3] for stroke length). 0 denotes zero width (a one-dimensional Euclidian line segment) and 1 denotes a stroke that is as broad as the typography's x-height (values above 1 are possible but uncommon – they express strokes that are broader than the x-height).

For each stroke in each letter, two measures are taken, one for its widest and one for its narrowest point. These are related to the x-height of the typography and then averaged in order to arrive at the WSR for the stroke.

$$\frac{\left(\frac{widest\ stroke\ width}{x-height} + \frac{narrowest\ stroke\ width}{x-height}\right)}{number\ of\ measures} \tag{1}$$

As illustrated in the middle top in Figure 4, the widest point of the stroke in 'o' is 17 pixels, the narrowest is 12 pixels. The x-height of the typeface is 46 pixels. Thus, the WSR of the heaviest section of the stroke in 'o' is 0.37 and 0.26 for the lightest section. The average WSR for the letter 'o' is 0.32. In our analysis of the entire logotype, we average WSR over all strokes in all letters in the logotype.

3.4.2 Expansion

In order to determine the relative width of letters in relation to their height, we measure the width of all letters and relate them to the x-height (or cap height in case of all caps): $\frac{letter\ width}{x-height}$. The 'o' of Ogol is 53 pixels wide and the x-height, again, is 46 pixels. Thus, the expansion measure is 1.15. Measures above 1 means letters are broader than they are wide as in the Ogol logotype. Again, we average over all letters in the logotype.

[3] This is to normalize a number of inequalities in our dataset: (1) Strokes are of unequal length, (2) logos are of unequal proportions, and (3) the bitmap images that serve as data are of unequal resolution. Therefore, we relate width measures to a constant that reasonably captures the scale of the logotype in a way that can be compared across all logos. We are aware that this lowers the external validity of our methodology by introducing a dependency on writing systems with an x-height and ascender/descender-like structure but, because all logotypes in the dataset are written in Latin letters, this does not harm the internal validity of the present study. However, in the interest of the scalability of the methodology, future research should attempt to overcome this bias.

3.4.3 Slope

In order to determine the slope of letters, a measure of the angle between a letter stem (or an approximation thereof) and the logotype's baseline is made. The output from Adobe Illustrator is formatted in degrees. 90° means that letters have no slope. 90–180° means that letters lean to the left, 0–90° means they lean to the right. We average the slope of all letters over the number of letters in the logotype. In the Ogol logotype, the average slope is 81.0.

3.4.4 Connectivity

Finally, in order to operationalize connectivity, the number of letter connections is related to the number of spaces between letters: $\frac{letter\ connections}{letter\ shapes}$. The Ogol logotype has 3 letter spaces. Of these, two are connected. The logotype's connectivity measure is 0.66.

3.5 Principal Component Analysis (PCA)

Applying the analytical framework introduced above provides 14 measurements that describe the shape, color, and typography of each logo in the corpus as summarized in Table 1. These measurements constitute the variables whose values reflect stylistic choices in logo design. However, comparing differences between individual variables is not likely to reveal distinctive stylistic preferences between oil industry and NGO logos, because they capture specific aspects of logos, which are assumed to form a tightly-integrated multimodal ensemble, as described in Section 2.

To understand how the 14 variables interact with each other, we therefore used Principal Component Analysis (PCA). PCA is a well-understood statistical method for multivariate analysis and dimensionality reduction (see e.g., Hervea & Williams 2010). Multivariate analysis, which seeks to describe the relation between multiple variables, has been productively applied in linguistics, perhaps most famously in the seminal work of Biber (1988), who used multivariate analysis to relate grammatical features to particular linguistic registers. Dimensionality reduction, in turn, refers to the process of reducing the number of variables (or 'dimensions' of variation) while preserving as much of the information contained in the original variables as possible.

PCA examines to what extent variables can be grouped by virtue of exhibiting similar patterns of variation so that they can be replaced by new variables, known

Tab. 1: Summary of the 14 variables measured for each logo.

Shape variables	
D/r icotype	The density of the icotype (count of shapes instances) averaged over count of regions in the icotype
S/D icotype	The proportion of straight shape features to the density in the icotype
A/D icotype	The proportion of angular shape features to the density in the icotype
C/D icotype	The proportion of curved shape features to the density in the icotype
D/r logotype	The density of the logotype (count of shapes instances) averaged over count of regions in the logotype
S/D logotype	The proportion of straight shape features to the density in the logotype
A/D logotype	The proportion of angular shape features to the density in the logotype
C/D logotype	The proportion of curved shape features to the density in the logotype
Color variables	
saturation	The average saturation of all colors in the logo
brightness	The average brightness of all colors in the logo
Typographic variables	
weight	The average WSR for all strokes in the logotype
expansion	The average expansion for all letters in the logotype
slope	The average slope of all letters in the logotype
connectivity	The proportion of connected spaces between letters to the total number of letter spaces in the logotype

as principal components. However, the principal components can never cover all the variation in the original data due to lower dimensionality, which leads to reduced capacity for explaining variation in the original data. For this reason, PCA analyses typically report how much of the original variation is explained by the principal components.

What makes PCA a particularly useful method is its ability to condense information while maintaining a mapping between the original variables and the principal components derived from them. This allows establishing how much each original variable contributes to each principal component. Each principal component can be thus interpreted in terms of the original measurements and related back to the analytical framework. In the field of multimodality research, PCA has been recently applied in a diachronic study of page layout in comics by Bateman et al. (2019), who use PCA to reduce 52 variables describing page layout to just five principal components, a much more manageable number of variables that nevertheless cover 61% of variation in the original data. An analysis of the principal components revealed temporal changes in the page layout along dimensions that would not have been discernible otherwise.

By condensing information about the original variables, the principal components can be seen as establishing the dimensions of variation in the corpus. As Bateman et al. (2019: 10) observe, comparing these dimensions against each other enables taking a topological perspective on variation. Previous accounts of linguistic genre have found such a perspective particularly useful, because from this perspective, the data may show differences and similarities along distinct dimensions of variation (cf. Lemke 1999). We consider this particularly beneficial for studying style in NGO and oil industry logos, which are likely to exhibit certain similarities by virtue of being logos, but also differences due to their respective domains.

We used the Python 3.8 programming language to perform PCA, using the implementation in the *scikit-learn* 0.23.2 library (Pedregosa et al. 2011). To preprocess the data, we used the *NumPy* 1.19.1 (Harris et al. 2020) and *pandas* 1.1.0 (McKinney 2010) libraries and performed statistical testing using the *SciPy* 1.5.2 library (Virtanen et al. 2020). The results were visualized using the *matplotlib* 3.3.1 (Hunter 2007) and *seaborn* 0.10.1 libraries.

4 Data Description

As noted above, the study is based on 50 logos, 25 each from the oil industry and non-governmental environmental organizations. The data collection began as a qualitative case study of a single, now decommissioned, logo for one Irish fuel and retail company called Topaz (Johannessen 2016) and grew from there: first as a comparison with other Irish fuel and retail brands in order to see if the original logo was typical of the forecourt and convenience retail business, then as a comparison between the downstream (consumer-facing) fuel and retail sector and the upstream (industry-facing) oil industry to see if the oil industry at large had a distinctive look and feel. Finally, the study evolved into a comparison of logos from the oil sectors and environmental NGO sectors.

The 25 oil company logos were selected according to the following criteria: 10 companies were taken from a list of the highest grossing fuel and convenience retail companies in Ireland in 2015 and 15 were taken from a Forbes list of the 25 highest grossing upstream oil companies in the World. The 25 environmental NGO logos were picked from an alphabetized list on Wikipedia. The list was found through a Google search on the search phrase “non-governmental organizations”. 25 entries were chosen by selecting every second logo on the list. Although there was no geographical rationale behind the sampling strategy, future research in the area might include such parameters as a possibly interesting dimension of variation. The full selection can be found in Table 3 in the appendix.

Figure 5 shows a small extract from of our data, three oil company logos (for Rosneft, PetroChina, and Royal Dutch Shell) and three environmental NGOs (Friends of The Earth, African Wildlife Foundation, and International Analog Forestry Network). The pictures were taken during the shape analysis phase, which was performed with the count tool in Photoshop. The original images of the 50 logos in the corpus, the measurements provided by applying our analytical framework and the code needed to reproduce the results reported below are available at: https://doi.org/10.5281/zenodo.3987747

5 Results

In order to explore possible *generic* differences in style between NGO and oil industry logos, we wanted to remove logos with rare stylistic features from the corpus. To achieve this, we applied a procedure known as z-score standardization to the measurements for each original variable. This procedure standardizes the values for each variable to have a mean value of zero and a standard deviation of one. As Bateman & Hiippala (2020: 4–5) explain, standard deviation measures how much a given value deviates from the mean value of a variable. The resulting Z-scores, which correspond to standard deviations, can be compared to a normal distribution to determine how far a value lies from the mean value for the given variable.

Normally distributed values form a bell-shaped curve which has useful properties for filtering the data for rare features: we know that for each normalized variable (out of the 14 in total), roughly 68% of the values are within one standard deviation on either side of the mean, that is, between values −1 and +1 (Bateman & Hiippala 2020: 4–5). We also know that 99.9% of values in a normal distribution fall within four standard deviations on either side of the mean. Conversely, the probability of finding a value below −4 or above +4 in a normal distribution is 0.01% or extremely rare. To remove logos with rare stylistic features, we excluded a logo if any of its variables had a value below −4 or above +4. A total of five such logos were removed from the corpus, retaining 22 logos for the oil industry and 23 for NGOs.

We then applied PCA to the standardized variables to reduce the number of dimensions from 14 to just five. The five principal components derived from the 14 variables cover 76.4% of the variation in the original data (see Figure 6).

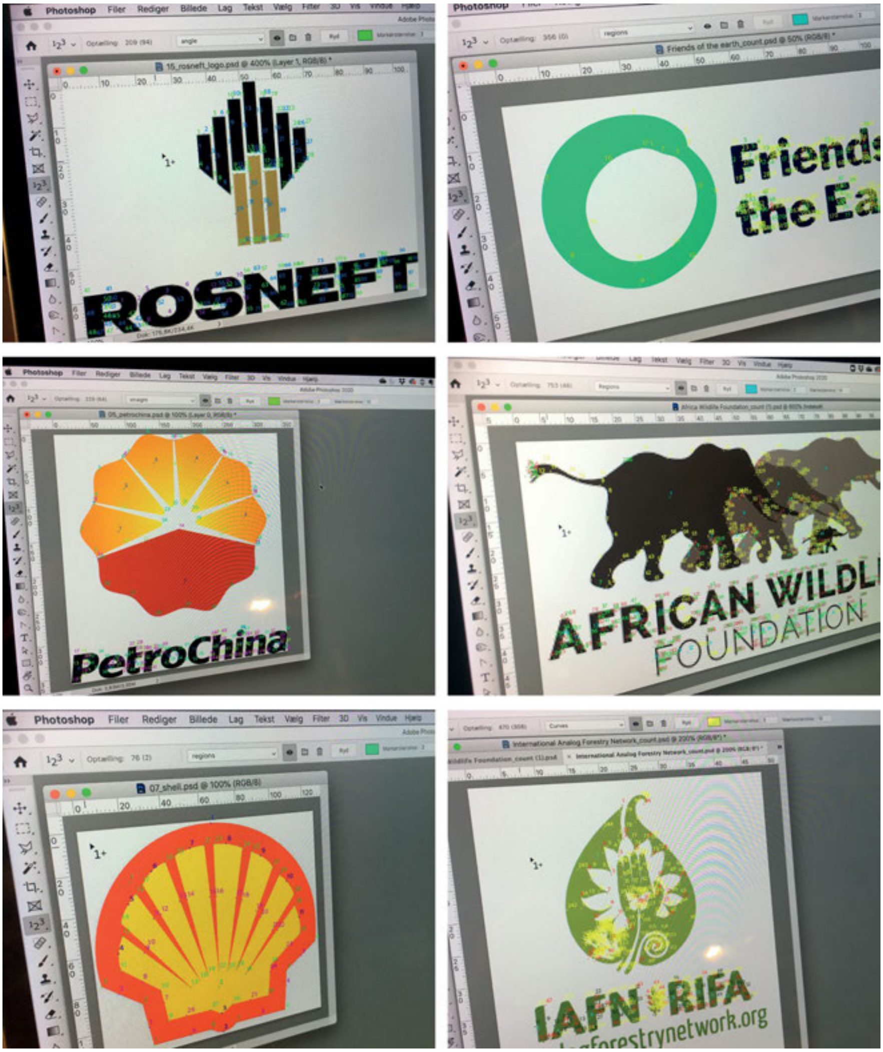


Fig. 5: Illustration of a selection of the data. In the left column are three oil company logos for Rosneft, PetroChina, and Royal Dutch Shell. In the right column are three environmental NGO logos for Friends of the Earth, African Wildlife Foundation, and International Analog Forestry Network. Pictures were taken during the shape analysis, which used the count tool in Adobe Photoshop.

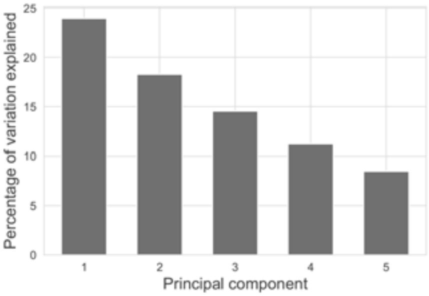


Fig. 6: Percentage of variation explained by the five principal components (total variation covered: 76.4%).

Tab. 2: Correlations between the original variables and principal components. The top-3 positive and negative correlates for each principal component are marked in bold and cursive, respectively.

	PC1	PC2	PC3	PC4	PC5
D/r icotype	*-0.394*	-0.235	**0.194**	**0.775**	*-0.472*
S/D icotype	**0.736**	**0.439**	-0.033	0.022	*-0.422*
A/D icotype	**0.698**	*-0.301*	-0.092	0.181	0.132
C/D icotype	*-0.865*	-0.09	0.018	0.046	**0.199**
D/r logotype	0.047	-0.03	0.016	0.105	-0.014
S/D logotype	0.193	*-0.319*	0.077	0.167	**0.24**
A/D logotype	0.29	*-0.707*	**0.108**	*-0.034*	-0.025
C/D logotype	-0.32	**0.696**	*-0.122*	*-0.067*	-0.113
saturation	0.161	0.113	*-0.809*	**0.313**	0.014
brightness	-0.186	-0.185	*-0.725*	0.076	0.096
weight	**0.193**	**0.485**	**0.372**	**0.487**	**0.409**
expansion	*-0.279*	-0.223	-0.057	*-0.098*	*-0.251*
slope	-0.013	-0.034	0.016	-0.008	-0.048
connectivity	-0.015	0.123	-0.12	0.253	0.236

The first principal component covers the most variation, because PCA seeks to maximize information about variation in each principal component while also avoiding correlations between principal components so that they do not increase or decrease together, but rather capture complementary aspects of variation in the original data. In other words, PCA attempts to describe variation in the original data as best as it can, while preserving as much information about variation as possible within a lower number of variables, that is, the principal components.

As explained in Section 3.5, each principal component can be mapped back to the original variables. This is achieved by examining the correlations between the original variables and the principal components. This is shown in Table 2, which shows how the original variables (rows) correlate with the five principal components (columns). These values, which run from −1 for negative correlation to +1 for positive correlation, while 0 indicates no correlation. For instance, we can observe that the first principal component PC1 is positively correlated with the proportion of straight shapes (S/D: 0.736) and angles (A/D: 0.698) in icotypes. These values are used as weights to determine how much each original variable contributes to the principal component in question when calculating its value.

Bateman et al. (2019: 11) call for attention to the fact that the principal components are simply a mathematical ‘best fit’ without regard to their possible meanings. Interpreting what each principal component *means* in the light of the annotation framework and the measurements made requires an additional step of analysis. We therefore summarize the positive and negative contributions of each original variable to the newly derived five principal components below:

- PC1 is positively correlated with straight and angular shapes in the icotype. It is negatively correlated with the density and proportion of curved shapes in the icotype, and to a lesser degree with the expansion of letters of the logotype.
- PC2 is positively correlated with curved shapes and the weight of strokes in the logotype, and negatively correlated with straight and angular shapes in the logotype, and to some extent with angular shapes in the icotype as well.
- PC3 is negatively correlated with saturation and brightness, and to some extent positively correlated with logotype weight.
- PC4 is strongly associated with the density of the icotype and to a lesser degree with the weight of the logotype and saturation.
- PC5 is negatively correlated with the density of the icotype and their proportion of straight shapes, as well as the expansion of the logotype. There is a slight positive correlation with the weight of the logotype and their proportion of straight shapes.

As pointed out in Section 3.5, these principal components establish the dimensions of variation in the corpus of logos, which may be observed from a topological perspective to identify similarities and dissimilarities between them. One way to examine the data from a topological perspective is to plot the dimensions of variation against each other for visual exploration, in order to generate hypotheses for statistical testing (O’Halloran et al. 2018: 20). Alternatively, the resulting visualizations may be considered as models, whose properties allow generating new knowledge about the phenomena under analysis (see Bateman & Hiippala, this volume).

To this end, Figure 7 plots the first principal component (PC1) on the horizontal axis against the other four principal components (PC2–5) on the vertical axis. In each plot, the colored dots represent individual logos, whereas the oval shapes stand for confidence ellipses, which are introduced in the caption of Figure 7. Note that the only purpose of coloring the dots is to distinguish between logos from NGOs (orange) and oil companies (blue) in Figure 7. Whether a given logo belongs to an NGO or an oil company is not used when calculating the principal components: any groupings observed emerge from the data and are not imposed by possible preconceptions about differences between the logos.

PC1 shows that oil industry logos prefer straight and angular shapes for the icotype, as indicated by their tendency to appear on the right-hand side of the plot (M=0.57, SD=1.32). NGO logos, in turn, favor curved shapes in the icotype, as reflected by their orientation towards the left (M=-0.55, SD=1.54). Here the extreme examples include the Friends of the Earth NGO logo at (-2.9, 3) and the Rosneft oil company logo at (3.4, 0.6) (for examples of actual logos, see the top row in Figure 5). Because a visual inspection shows considerable overlap between the two samples in the middle region, the possible difference must be evaluated statistically: the difference between oil industry and NGO logos along PC1 was found to be statistically significant with a nearly large effect size (Mann-Whitney U=152.0, p=0.01; Cohen's d=0.73).[4]

For PC2, which is plotted against PC1 in the top-left hand corner of Figure 7, the plot reveals that oil industry logotypes seem to be characterized by curved shapes and weight, as indicated by the positive values and a smaller, upward leaning confidence ellipse (M=0.14, SD=1.06). These features are also characteristic of two NGO logos in the top-left corner of the plot. Generally, NGO logos show greater variation, as indicated by the larger confidence ellipse (M=-0.14, SD=1.52), but appear to prefer straight and angular shapes for the logotype, as reflected by their negative values for PC2. However, this difference is not statistically significant at $p < 0.05$ (Mann-Whitney U=190.0, p=0.078; Cohen's d=0.44).

Along the dimension of variation represented by PC3, which is strongly negatively correlated with saturation and brightness, shown on the top-right hand side of Figure 7, the oil industry logos show greater variation (M=-0.26, SD=1.38), but are generally characterized by bright and saturated colors. These features seem to be less prominent for NGO logos (M=0.25, SD=0.94), which suggest that they prefer darker and non-saturated colors. The difference along PC3 is statistically significant with a medium effect size (Mann-Whitney U=165.0, p=0.02; Cohen's d=0.62).

[4] We used the Mann-Whitney U test for significance testing, because the samples are not normally distributed. For more information, see Bateman & Hiippala (2020: 8–10).

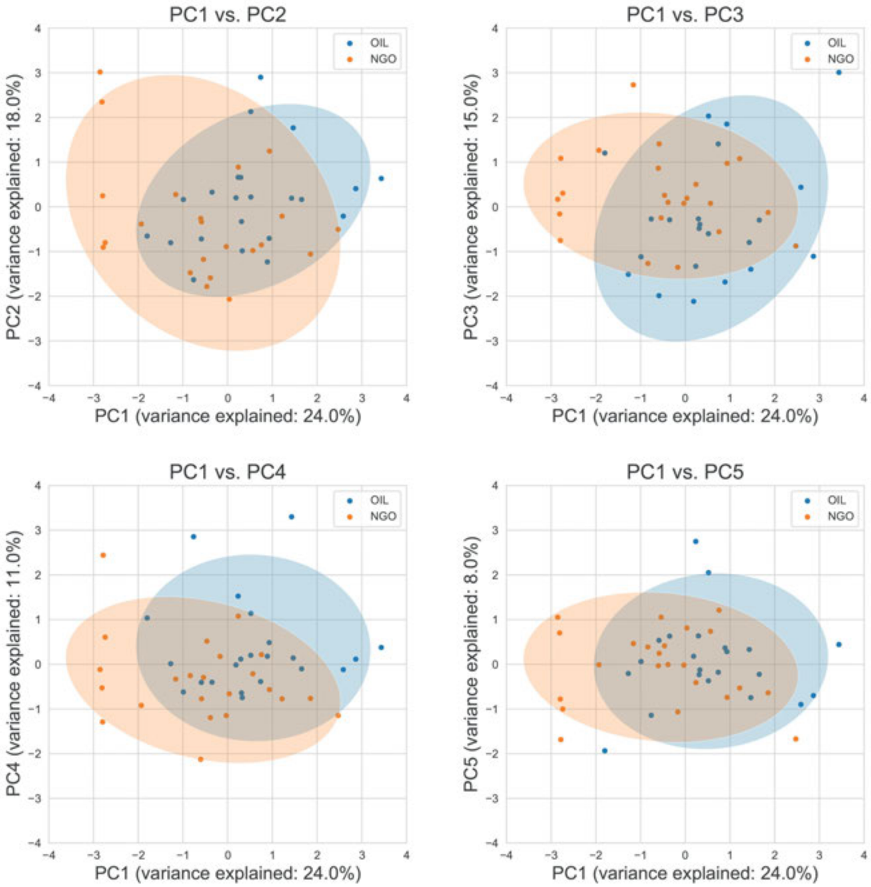


Fig. 7: Plotting the principal components against each other. The area covered by each oval-shaped confidence ellipse corresponds to two standard deviations from the mean value, which covers 95% of the data in a normal distribution. If the analysis were re-run a number of times, the mean value for a given logo type would fall within its confidence ellipse with 95% probability. The size of a confidence ellipse, in turn, is an indication of how much the samples deviate from their mean value.

PC4 and PC5, which are plotted against PC1 in the lower part of Figure 7, account for 11% and 8% of the variation, respectively. For PC4, which is positively correlated with icotype density, logotype weight and saturation, the samples show a statistically significant difference with a large effect size (Mann-Whitney U=138.0, p=< 0.01; Cohen’s d=0.845). Comparing the original measurements captured by PC4 revealed that the difference results from saturation (U=118.5, p=< 0.01; d=1.02), not logotype weight (U=247.0, p=0.45; d=0.04) nor icotype density (U=191.0, p=0.08; d=0.43). The difference along PC5 was not statistically significant at Mann-Whitney U=252.0, p=0.5; Cohen’s d=0.007.

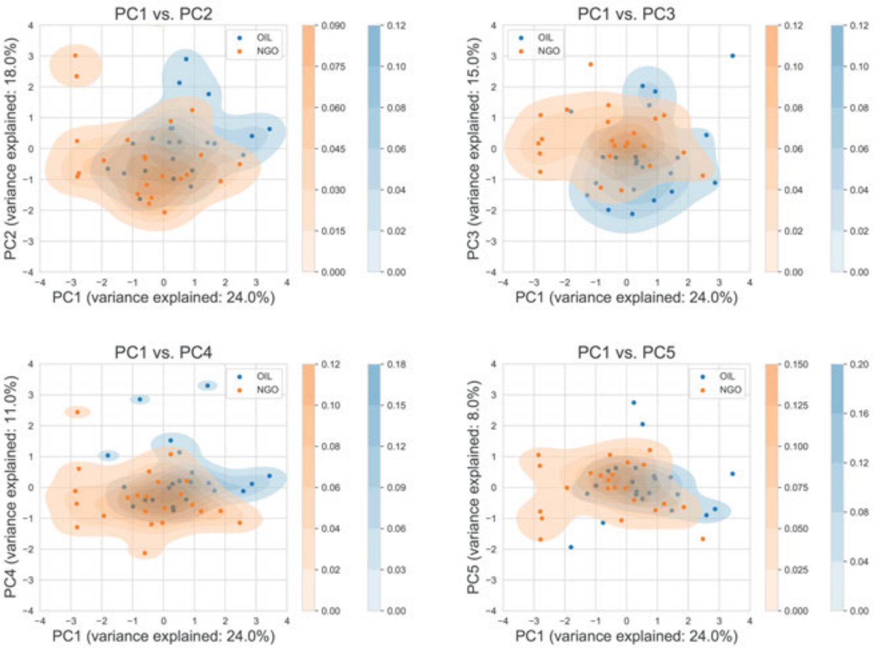


Fig. 8: Kernel density estimations for the principal components.

While the confidence ellipses in Figure 7 help to establish how each sample varies and along which dimensions, they cannot reveal whether logos are more likely to occur within a given region of the plot. To investigate how individual logos are distributed over particular regions, Figure 8 applies kernel density estimation (KDE) to the principal components. KDE estimates the likelihood of finding a logo of a given type within a particular region of the plot. This likelihood is measured using a probability density function, whose values are mapped to particular shades by the bars on the right-hand side. Note that the lowest interval at the bottom of the scale is not colored to make the regions with higher probability stand out in the plot.

Figure 8 reveals that along the dimension of variation represented by PC2 in the top left-hand corner, the NGO logos form two distinct regions. Whereas most NGO logos are likely to be found in the middle region, the upper group is consistent enough to form its own cluster. The upper group, which is characterized by weighted and curved shapes in the logotype, may represent a trend that diverges from the majority of NGO logos. Whether this represents an emergent trend or the sample is too small to bridge the gap between these two groups would warrant further research.

What is notable across all plots in Figure 8 is that the regions associated with the highest probability densities for both oil and NGO logos are close to each other. These regions are likely to capture the common features of logos, that is, shapes and colors that allow us to recognize these artifacts as logos in the first place. It is also worth noting that oil industry logos along PC4 and PC5 seem to exhibit less variation than their NGO counterparts, as reflected by their smaller area and higher values for probability density function. In other words, the oil industry logos are less flexible in terms of their choices pertaining to shape and color.

Finally, we now summarize the main findings in Figures 7 and 8 separately for both oil industry and NGO logos. For the oil industry logos, the results suggest that:

- Oil industry logos prefer straight and angular shapes in the icotype (PC1).
- Oil industry logos favor curved shapes and weighted strokes in the logotype (PC2).
- Oil industry logos prefer bright and saturated colors (PC3).
- Oil industry logos are less varied.

For the NGO logos, the results suggest the following:

- NGO logos prefer curved shapes for icotypes (PC1).
- NGO logos favor straight and angular shapes and lighter strokes in the logotype (PC2).
- NGO logos prefer darker and non-saturated colors (PC3).
- NGO logos are more varied.

6 General Discussion

Our study shows that the logos of oil companies and environmental NGOs are subtly but reliably different. We have found statistically significant differences on PC1 (associated with shape distributions in the icotype), PC3 (associated with color), and PC4 (associated with the weight of the logotype, the density of the icotype and color saturation). Therefore, on the gloss of style we introduced in Section 2, which highlights “manner of expression” over “matter expressed”, we argue that we have indeed observed an interesting difference between the look of oil company and environmental NGO logos. Because the two groups of logos are associated with different industrial sectors, one for profit, the other non-profit, the study lends empirical weight to our initial idea: That the logos in organizations' visual identities express something akin to van Leeuwen's notion of "social style". This could in this case be understood as expressions of social allegiance to an industry or business sector.

We have designed the study to emphasize "manner of expression" by looking only at formal variables of shape, color, and typography, and to de-emphasize everything to do with the "matter expressed" by those forms. Our study says nothing about preferences in the two groups for, for example, icotypes that are fairly abstract or specify animals and plants, or about logotypes with preferences for shorter or longer host names. Whether the formal difference we have observed in fact amounts to a difference in people's experience of "look and feel" of these logos remains an open question, which we would have to pursue within a framework of phenomenology or reception analysis. However, experimental findings already make us expect that such differences will in fact be experienced differently. In a 2015 study of crossmodal[5] affective correspondences in the relationship between typeface and taste (Velasco et al. 2015) the authors found participants to strongly associate round shapes in typefaces with sweetness (and liking) and angular shapes with bitter, salty and sour (and dislike). The authors conclude that "it is possible to hypothesize that those who associate curved typefaces with more positively valenced emotions will be more likely to associate them with sweetness, a taste that is known for its positive hedonic effect on humans" (2015: 8). Although we have looked at graphics in a very different context, we would also expect to see differences in experience arising from differences in form.

Many things in the world are built of graphic shape, color, and typography. The analytical and annotational framework used in this explorative and data driven study is generic enough to equally describe logos, smart watch interfaces and public signage. Readers may find it a weakness of the study that the logos in the two groups cluster so closely around the same probability density regions. However, we find the statistical significance of the differences we have found to be very encouraging. Had we chosen to compare with logos for martial arts studios, street wear brands, or toy companies it seems very possible that we would have found even more significant differences.

[5] Velasco et al. are experimental psychologists and use the term modality to refer to sensory modalities.

We assume that, in order to satisfy people's expectations for what a logo should look like in order to distinguish it from a road sign or smart watch interface, certain criteria must be met. Maybe the fact that so many of the logos in our two samples cluster around the same probability density regions is a result of the way their designers seek to balance the simplicity, which the graphic design profession holds in such high regard, with the need for distinctive brand identity (smart watch interfaces are more complex and road signs are less distinctive). They cluster together because they look like logos. If such an assertion holds, it is interesting that there is greater overall variability in the NGO sub-sample than in the oil sub-sample (as indicated by the size of the 95% confidence ellipses). We can only speculate why, but possibly grass roots organizations operate under budget constraints that cause them to use logos made by more amateur designers with less of an instinctive feel for what is called for in order to build strong branding.

We are very aware that our analytical scheme (even if it has in fact produced a robust difference between the two groups) is not yet developed to a grain that would suffice to capture more subtle aspects of the look of many middle-of-the-road logos. We have not looked at spatial distribution of regions, shapes, and colors. For example, a simple proportion between curvature and angularity says nothing about the regularity of shape. A region made up of 100 identical and evenly spaced curves looks considerably different than one made up of a 100 unevenly spaced and different curves. Similarly, it does not account for the relative salience of shape features. A square with smooth corners will look like an uneven circle if the curves have a big enough radius. Such features are very likely to factor into people's experience of shape – but making them operationally viable for manual annotation is a challenge we have yet to overcome.

The way we have framed this argument does not warrant too much speculation about the values and ideologies that regulate the oil industry and environmental altruism respectively. We do, however, find that our results are sufficiently robust to justify a qualitative follow-up study of consumers' and designers' receptions or experiences seeking to connect our findings with the common notion that "brands are people" and so with social values of, e.g., nature vs. culture, with aesthetics of softness vs. hardness, with ideologies of humanism vs. corporatism, to name a few.

Acknowledgements and Funding

Christian Mosbæk Johannessen acknowledges the Innovation Fund Denmark's Industrial Ph.D. programme [grant no. 0604-03422] and Danfoss A/S as having funded part of the research reported here.

Appendix: Overview of Logos

Tab. 3: Alphabetical overview by industry of the fifty logos in the sample. The six logos shown in Figure 5 are marked in cursive.

Oil logos	**Environmental NGO logos**
Abu Dhabi National Oil	350.org
Applegreen	Arab Forum for Environment and Development
ASDA	*African Wildlife Foundation*
Atlantic Petroleum	American Forests
Chevron	*Analog Forestry Network*
Eesti Energia	Biodiversity International
Eni	Center for Development and Strategy
Gazprom	Climate Action Network
GreatGas	Conservation International
Iranian Oil	Deep Green Resistance
Kuwait Petroleum Corporation	Earth Day Network
Maxol	Environmental Defense Fund
Morris Oil	Foundation for Environmental Education
National Iranian Oil Company	Frankfurt Zoological Society
OMV	*Friends of the Earth*
Pemex	Forest Stewardship Council
PetroChina	Global Witness
Petronas	Green Actors of West Africa
Quatar Petroleum	Greenpeace
Rosneft	Ideas for Us
Royal Dutch Shell	International Rivers
Solo	International Union for Conservation of Nature
Star	Mountain Wilderness International
Tesco	Oceana
Top	Pragya

Bibliography

Adams, S., N. Morioka & T. L. Stone. 2006. *Logo Design Workbook: A Hands-On Guide to Creating Logos*. Rockport Publishers.

Aiello, G. 2017. Losing to Gain: Balancing Style and Texture in the Starbucks Logo. In T. Johannessen, C. M. & van Leeuwen (ed.), *The Materiality of Writing. A Trace Making Perspective*, London: Routledge.

Airey, D. 2015. *Logo Design Love. A Guide to Creating Iconic Brand iIentities*. San Francisco, CA: New Riders 2nd edn.

Baines, Phil & A. Haslam. 2005. *Type & Typography*. London: Laurence King Publishing.

Bateman, J. A., F. O. D. Veloso & Y. L. Lau. 2019. On the Track of Visual Style: A Diachronic Study of Page Composition in Comics and its Functional Motivation. *Visual Communication* https://doi.org/10.1177/1470357219839101.

Bateman, John. A. 2008. *Multimodality and Genre. A Foundation for the Systematic Analysis of Multimodal Documents*. New York: Palgrave Macmillan.

Bateman, John A. & T. Hiippala. 2020. Statistics for Multimodality: Why, When, How – An Invitation. SocArXiv https://doi.org/10.31235/osf.io/7j3np.

Beebe, B. 2004. The Semiotic Analysis of Trademark Law. *UCLA Law Review* 51(3). 621–704.

Bertin, J. 1983. *Semiology of Graphics: Diagrams, Networks, Maps*. Madison: University of Wisconsin Press.

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Bishop, R. 2001. Stealing the Signs: A Semiotic Analysis of the Changing Nature of Professional Sports Logos. *Social Semiotics* 11(1). 23–41.

Cowin, E. & J. Matusitz. 2011. The Ongoing Transformation of the McDonald's Logo: A Semiotic Perspective. *Journal of Visual Literacy* 30(2). 20–39.

Gill, Eric. 1988. *An Essay on Typography*. Boston, MA: David R. Godine Publisher.

Harris, Charles R., K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. Fernandez del Reo, M. Wiebe, P. Peterson, P. Gerard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke & T. E. Oliphant. 2020. Array Programming with NumPy. *Nature* 585. 357–362. https://doi.org/10.1038/s41586-020-2649-2.

Heilbrunn, B. 1997. Representation and Legitimacy: A Semiotic Approach to the Logo. In W. Nöth (ed.), *Semiotics of the Media. State of the Art, Projects and Perspectives*. Berlin: De Gruyter Mouton.

Heilbrunn, B. 2001. *Le Logo*. Paris: Presses Universitaires de France.

Hervea, A. & L. J. Williams. 2010. Principal Component Analysis. *WIREs Computational Statistics* 2(4). 433–459. https://doi.org/10.1002/wics.101.

Hunter, John D. 2007. matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* 9(3). 90–95.

Hyndman, S. 2016. *Why Fonts Matter*. London: Random House.

Jewitt, C. 2013. Multimodal Methods for Researching Digital Technologies. In S. Price, C. Jewitt & B. Brown (eds.), *The Sage Handbook of Digital Technology Research*, London: Sage.

Johannessen, C. M. 2011. *The Forensic Analysis of Graphic Trademarks. A Multimodal Social Semiotic Approach*. Odense: University of Southern Denmark dissertation.

Johannessen, C. M. 2016. Experiential Meaning Potential in the Topaz Energy Logo: A Framework for Graphemic and Graphetic Analysis of Graphic Logo Design. *Social Semiotics* 1–20.

Johannessen, C. M. 2017. The Challenge of Simple Graphics for Multimodal Studies. Articulation and Time Scales in Fuel Retail Logos. *Visual Communication* 16(4). 163–185.

Koller, V. 2009. Brand Images: Multimodal Metaphor in Corporate Branding Messages. *Multimodal Metaphor* 11. 45–72.

Kress, G. & T. van Leeuwen. 2002. Colour as a Semiotic Mode: Notes for a Grammar of Colour. *Visual Communication* 1(3). 343–369.

van Leeuwen, T. 2005a. *Introducing Social Semiotics*. Abingdon: Routledge.

van Leeuwen, T. 2005b. Typographic Meaning. *Visual Communication* 4(2). 137–143.

van Leeuwen, T. 2006. Towards a Semiotics of Typography. *Information Design Journal* 14(2). 139–155.

Lemke, Jay L. 1999. Typology, Topology, Topography: Genre Semantics. Manuscript University of Michigan.

Llorente-Barroso, C. & F. Garciaa-Garciaa. 2015. The Rhetorical Construction of Corporate Logos. *Arte, Individuo y Sociedad* 27(2). 289–309.

Lupton, Ellen. 2014. *Thinking with Type: A Critical Guide for Designers, Writers, Editors, & Students*. New York: Princeton Architectural Press.

McKinney, Wes. 2010. Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (eds.), *Proceedings of the 9th Python in Science Conference*, 51–56.

Mollerup, P. 1997. *Marks of Excellence. The History and Taxonomy of Trademarks*. London: Phaidon.

O'Halloran, K. L., S. Tan, D. S. Pham, J. Bateman & A. Vande Moere. 2018. A Digital Mixed Methods Research Design: Integrating Multimodal Analysis with Data Mining and Information Visualization for Big Data Analytics. *Journal of Mixed Methods Research* 12(1). 11–30. https://doi.org/10.1177/1558689816651015.

Pedregosa, Fabian, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot & É. Duchesnay. 2011. scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12. 2825–2830.

Reimar, Eli & H. Birkvig. 2003. *Skriftanalyse*. Kobenhavn: Grafisk Litteratur.

Scott, D. 1993. Air France's Hippocampe and BOAC's Speedbird: The Semiotic Status of Logos. *French Cultural Studies* 4(11). 107–127.

Silva, L. E. V., A. C. S. S. Filho, V. P. S. Fazan, J. C. Felipe & L. O. M. Junior. 2016. Two-Dimensional Sample Entropy: Assessing Image Texture through Irregularity. *Biomedical Physics & Engineering Express* 2(4). https://doi.org/10.1088/2057-1976/2/4/045002.

Skaggs, S. 2017. *FireSigns. A Semiotic Theory for Graphic Design*. Cambridge, MA: The MIT Press.

Skaggs, S. 2018. Visual Identity: Systems and Semiotics. *The American Journal of Semiotics* 34(3/4). 313–330.

Stötzner, Andreas. 2003. Signography as a Subject in its own Right. *Visual Communication* 2(3). 285–302.

Velasco, C., A. Woods, S. Hyndman & C. Spence. 2015. The Taste of Typeface. *i-Perception* 6(4). 1–10.

Virtanen, Pauli, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt & SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17. 261–272. https://doi.org/10.1038/s41592-019-0686-2.

Wheeler, A. 2013. *Designing Brand Identity. An Essential Guide for the Whole Branding Team*, vol. 2. Hoboken, NJ: John Wiley and Sons.

Hartmut Stöckl

Pixel Surgery and the Doctored Image

The Rhetorical Potential of Visual Compositing in Print Advertising

Abstract: Based on a corpus of 232 print advertisements, the chapter studies the function of the composited, i.e., computer-generated, image (CGI) for the construction of multimodal arguments. Corpus annotation first captures design operations as manipulations or configurations of visual structure. This forms the basis for an enquiry into the images' rhetorical potentials, that is their function for facilitating multimodal argumentation. Rhetorical potential is supplemented by annotating for argument type and relational propositions in the text-image relations. Besides producing an empirically verified, rhetorically motivated typology of print-CGI in advertising, the study also illustrates prominent and potent ways of building multimodal arguments through text-image relations.

Keywords: composited CGI, text-image relations, multimodal argumentation, design operation, rhetorical potential, relational proposition

1 Introduction: Motivating the Research

Graphic design for the advertising industry has always been at the forefront of innovations and keeps spearheading the implementation of new technologies. When, in the 1980s, computer-generated imagery (CGI) started to be used in filmmaking, it was soon adopted by advertisers in commercials, and all it took for CGI to be transferred to the making of static visual ad-images was increased computing powers and decreasing costs. Such images — which have loosely been called doctored imagery and also go by such labels as 'digital imaging/art' (Lürzer's International Archive 2013/2014: 12–14) or 'photo-compositing' (Bühler et al. 2017: 76–95) — form the focal point of departure for the multimodal research presented here. They establish an image type that seemingly presents photographic reality but actually contains all kinds of strategic design manipulations concerning individual visual image elements (VIEs), visual styles, pictorial logic, and rhetorical intent.

Recording current trends in avantgarde advertising and graphic design benchmarks, *Lürzer's Archive* acknowledged the move towards digital art by publishing a first collection of *200 Best Digital Artists Worldwide* in 2013. The editors referred to the contents as "print CGI [...], [a] hybrid form of imaging, which blurs the

https://doi.org/10.1515/9783110725001-008

borders between photography and illustration" (Lürzer's International Archive 2013/2014: 11). Aware of the difficulties in defining this novel image type, they went on to stipulate that "photos that looked like they were photos, and illustrations that looked as if they had been done using analogue technology were not what we wanted to show in this book" (Lürzer's International Archive 2013/2014: 11). In this sense, digital art is not the same as merely digitally retouching a photograph, it rather refers to "a new realistic aesthetic which is not photography" and is mainly enabled by "creating images from 3D-programs" (Lürzer's International Archive 2019/2020: 8–11). The main argument that is made about print CGI is that it "has a lot more flexibility to support more varied ideas" (Lürzer's International Archive 2019/2020: 8–11), and that it has led or is leading to "the explosion of new ideas" (Lürzer's International Archive 2013/2014: 12–14).

This claim about a novel image-making freedom sparked my multimodal interests and provoked three basic questions that led my research design:

1. Based on a sufficiently large corpus of such CGI-ad-images, how do we ascertain the *design operations*, i.e., the kinds of manipulations of visual structure, style, and content that underlie distinct types of digital art? Can a *classification system* or some kind of typology be derived from these empirical enquiries into underlying design operations that could prove useful in ordering a seemingly diverse and varied image terrain?
2. With such a typology of design operations in place, how do we determine the *rhetorical potential* of each class or group of visual manipulations? Rhetorical potential is to be understood here as the suitability or powers of an image to play a functional role in supporting a *multimodal argumentation*, i.e., a text-image relation calculated to express some entrenched and conventionalized pattern of argumentation, such as exemplifying, comparing, or causal conclusions, etc.
3. Assuming text and image link in specific ways to facilitate a persuasive argument, how do we make statements about the *relational propositions* that each type of design operation promotes? In other words, how do text and image interrelate in order to fabricate a *logical-semantic relation* between the contents expressed visually and verbally? Here it would be worth exploring questions of the relative status of text and image and the kinds of inter-semiotic cohesion deployed in order to make relational propositions and arguments work.

2 Theoretical Background

The present study connects to and feeds back into essentially three broad, interconnected areas of enquiry: pictorial theory with an eye to image classification, visual rhetoric, and multimodal argumentation. I will set these out very briefly here and show how advertising is an ideal testing ground for the general questions raised in these fields of study.

2.1 Pictorial Theory

Digital art or print-CGI poses interesting and practically relevant questions for image classification. In theories of visual communication (cf. Müller & Geise 2015: 211–217) and in pictorial theory (cf. Stöckl 2004: 115–126), it is a well-established idea that images can be classified in manifold ways, e.g., in terms of their content, depictional/referential strategies, medial/material and technological properties, or rhetorical techniques, etc. Such typological endeavors are not purely theory-oriented, they also become relevant in the practice of archiving visual/multimodal material as, for instance, in the advertising industry. *Lürzer's Archives*, for one, have been confronted with issues of how to sort the selected digital art (cf. Lürzer's International Archive 2013/2014, 2019/2020) into meaningful categories so that readers can search for them. They opted for a mixed-categories approach, which adopts the following criteria:

1. content (e.g., *animals*, *landscape*, *fashion*, *people*, *work*)
2. product type (e.g., *automotive*, *food/drink*, *industrial*, *transportation*)
3. relation to reality (e.g., *fantasy worlds*)
4. formal/structural/material aspects (e.g., *character design*, *typography*, *still life*, *objects*)

Other options for classification would be content alone (cf. Müller & Geise 2015: 211) or assumed pictorial functions in a given genre (for advertising cf. Stöckl 2009: 14). The rationale adopted in the present study is to base the typology on what I decided to call *design operations*, i.e., more or less clearly discernible ways of manipulating or structuring visual image elements and entire images — such as combining, substituting, juxtaposing or modeling — that are facilitated by digital image-making technologies. Visual image elements correspond to any easily recognizable pictorial gestalt that clearly refers to an object, situation, or action. In this fashion, material-technological properties of digital art are put centre-stage and form the point of departure for observations on rhetoric, multimodal texture, and argumentative structure.

This rationale acknowledges the notion that digital imagery is a new medium different from traditional (and retouched) photography and manual illustration, which engenders specific ways of meaning-making and has its distinct multimodal and rhetorical potentials. There is, in the advertising profession, a clear awareness of the typical affordances of and pragmatic differences between largely un-doctored photography and digital image-making. This reflects in a knowledge about which image type is suitable for which visual content or persuasive idea. Lifestyle, fashion, and food, for instance, largely necessitate un-doctored photography, whereas print-CGI is primarily suited for fictitious content or referents that are hard to find or get to and impossible to model in traditional ways, but also for interiors and visual metaphor (cf. Lürzer's International Archive 2013/2014: 12–14). Generally, digital art raises interesting semiotic issues of reference, i.e., the relation between the signs used and the objects they refer to. This relation could also be termed 'depictional mode' of an image (Stöckl 2004: 115–116) and would allow such plausible and common distinctions between denotational, fictitious, and non-denotational images (Scholz 1998), but also between illusionary, exemplifying, or allegorical images. From my corpus observations, three quite distinct 'depictional modes' are common:

1. *Recognizable fiction*: Images present fictitious worlds, which may be possible and impossible illusionary scenarios that can only be realized thanks to CGI.
2. *Seeming reality*: Images present a seamless recreation of reality in order to insinuate indexical photography, which often seems hyper-real.
3. *Emulated pictorial/media-styles*: Images imitate the material feel of media other than photography (e.g., postcard, packaging), irrespective of the objects/referents shown.

'Seamlessness' appears to be an interesting concept in relation to print CGI, which entails that the design operations are maximally invisible to the naked, untrained eye so that the image — while being the result of many fragments in various stages of production — seems a holistic entity. For the collation of a corpus of digital advertising art (see Section 3 below), seamlessness poses problems that can only be overcome by informed scrutiny, enquiries with the makers, and production info provided in the archives.

2.2 Visual Rhetoric

Despite understandable misgivings about rhetoric for not “lend[ing] itself to precise descriptions” (Bateman 2014: 132), rhetorical accounts of visual communication have been quite prolific and have proved fruitful for multimodal approaches (cf. Bateman 2014: 119–136). Although not limited to it (cf. Stöckl 2014a), such approaches have focused mainly on a transposition of rhetorical figures to the image or to text-image combinations and their classification (e.g., Gaede 1981; Durand 1987; McQuarrie & Mick 1996; Phillips & McQuarrie 2004). In these endeavors, the notion of ‘rhetorical figure’ has been conceived of as a deviant and therefore salient configuration of formal and content elements (i.e., propositions). Deviation in turn provokes an involvement of the recipient that is higher than usual, as norm-based expectations are being disappointed. It is, therefore, plausible that scholars have been adamant to distinguish degrees of deviance which they sought to connect with cognitive activation, recognizability, and ease of understanding (cf. e.g., McQuarrie & Philipps 2005).

Any genre-based account, however, will have to acknowledge that rhetorical figures may perhaps more suitably be seen as conventionalized cognitive strategies which come in useful to fulfill persuasive and argumentative functions required in a given type of communication. This is the view I endorse here, following Bateman’s general notion that “the delimitation of just what is a rhetorical device and what is not is essentially a functional one” (Bateman 2014: 122). In visual rhetorical studies, functionality has been interpreted differently, as the recognition of rhetorical deviance (Durand 1987), as salience or ease of recognition (McQuarrie & Mick 1996; Phillips & McQuarrie 2004), as audience effects (McQuarrie & Philipps 2005; van Mulken et al. 2005), or as specific communicative functions (Gaede 1981). In relation to a product-analytical study of advertisements, which need to substantiate a claim, it seems useful to define the function of rhetorical operations as their *rhetorical potential* (see Section 2.3) for argumentation or *persuasive usefulness*.

The approach in the present study does not work with established rhetorical figures as such, but rather with what classifications have termed ‘rhetorical operations’ (see Durand (1987: 295) and McQuarrie & Mick (1996: 426)) or ‘visual structure’ (Phillips & McQuarrie 2004). In this fashion, I focus on the one rhetorical criterion that deals with manipulations of graphic-semiotic material, i.e., visual elements that can be configured in the space of a digitally manipulated image. By singling these out, I hope to avoid the pitfalls of delineating concrete rhetorical figures, which are notoriously hard to ascribe to visual images or text-image combinations. This way, I also steer clear of an unnecessary complexity and reserve the issue of text-image relations or multimodal rhetoric for a treatment of the more general, coherence-oriented relational propositions (Mann & Thompson 1986; see Section 4.4).

2.3 Multimodal Argumentation

The genre of advertising is clearly of the argumentative type, which means it must — in whatever semiotic mode-combination or medium — make a claim about the product or brand in question and substantiate it by positive appraisal. McQuarrie & Philipps (2005: 7–9) state that indirect claims have become far more prominent since the 1990s and that images and rhetorical figures play a major role in construing them. Overall, approaches to multimodal argumentation in advertising seek to determine whether, to what extent, and exactly how images are instrumental in building arguments and which place rhetorical figures have in visual and multimodal argumentation.

Generally, studies of multimodal argumentation have gained much ground recently (Tseronis & Forceville 2017; Tseronis & Pollaroli 2018; van Eemeren & Garssen 2012). They have both addressed foundational issues as well as applications to diverse genres and media. Bateman (2018: 295) claims that "arguments can indeed be pursued multimodally" and makes two points that are relevant to the present study: First, it is discourse structures elaborated in a given mode that contribute to and construct arguments, not modes per se or media as such (Bateman 2018: 302). In this sense, both images and text can contribute propositional content that may enter into argument construction. Second, each semiotic mode brings its own specific affordances to bear on the argument (Bateman 2018: 302–303), resulting in essentially reciprocal mode elaboration. Images, in this respect, are usually seen to substantiate verbal claims, but could also perform different semiotic and rhetorical work. Kjeldsen (2012: 251–252) generally describes images as affording 'thick interpretation', i.e., pictorial meanings are multiple and varied; they are only tied down in context and will, in acts of inference, generate a number of recipient-specific weak or strong implicatures (cf. Forceville 2014).

Two essential questions can be raised about text-image argumentation: What is the relative status of the semiotic modes, and what is the role of rhetorical operations or figures in constructing the argument and giving it persuasive force? Regarding the first question, even though in print advertising, images are perceptually dominant and rhetorically potent, they need not necessarily have a higher relative status. According to my corpus observations, images are hardly ever 'visual flags' (Roque 2012: 281) merely attracting attention to and illustrating a verbal argument. Mostly, they rather have equal status and jointly (cf. joint argument Roque 2012: 283–284) construe a multimodal argument. Regarding the second question, I support the view (Kjeldsen 2012: 243–244) that rhetorical operations generally facilitate multimodal argument construction. Operating either in text or image or in both, rhetorical figures help recipients recognize discourse elements crucial to the argument and suggest familiar ways in which text-image relations may be read.

As suggested in Section 2.2, the present study adopts the rationale that rhetorical operations in the computer-generated/-manipulated image open up specific potentials for argumentation and enable various types of arguments as defined in rhetoric (Janich 2013: 131–136; Lehn 2011: 164–183).

3 Material and Method

The rationale of the present study, then, is to empirically determine the types of rhetorical operations (here labelled ‘design operations’), which are typically deployed in the computer-generated advertising image, and systematize these in a typology. In order to gauge the genre-specific functions of these design operations, I look at their potential for constructing multimodal arguments. In other words, I inspect the formal-graphic manipulations for what they offer to a cognitive pattern that can sustain a multimodally construed argument. This includes scrutinizing the types of arguments constructed and studying the relational propositions established between image and text.

3.1 Corpus

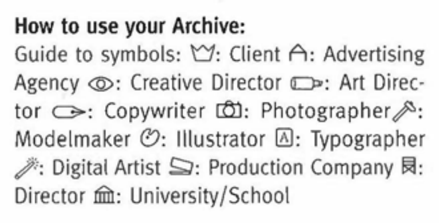
The first task the present study faced was to collate a suitable corpus that would assemble topical cases of digital art or composited images in the print advertising genre. To sample relevant data, I opted for *Lürzer’s Archive* because its editors collect particularly ingenious and graphically outstanding work of the advertising industry worldwide. While this source does not provide any indications of which specific media the ads were published in, it clearly represents global high-quality, often award-winning advertising, which covers the whole breadth of potential target audiences and product types (altogether 32 categories). The corpus used in the present study consists of a sampling of all ads in volumes 1–6 of 2019, yielding a total of 564 texts. These were subdivided into four groups designating material-technological types of image-making or graphic production: doctored images, un-doctored images, illustrations, and type only. The latter group will be ignored here as it does not feature images as such.

It is useful at this point to add some ideas about how the three image types differ and what their potential uses are in the advertising genre:

1. *Un-doctored images* are traditional photographs of people, objects and scenarios, but they may also be x-ray, MRI or other types of 'technical' pictures generated by image-making technologies. The semiotic essence of the un-doctored image is indexicality and iconicity, i.e., a more or less faithful representation of reality in all its physical detail. In advertising, such images prove the existence of entities of the surrounding world; they engage in showing their essence or optical beauty and, for this purpose, often employ simple retouching.
2. *Illustrations* set great store by visible materiality on the graphic surface, that is, they exhibit various styles of sketching, drawing, painting, etc. Illustration may tend towards a strong indexicality when its aim is to seem like a faithful depiction of visible reality, or it may aim to create fiction by giving real or imaginary scenarios the feel of the fabricated. Most useful to advertisers is illustration's ability to evoke associations with recognizable graphic styles that indicate social values and orientations that favorably connect with audiences and products.
3. *Doctored images* involve a large variety of more or less conspicuous manipulations of image elements and graphic form that have strong repercussions on pictorial content and function. Rather than offer truly indexical or photographic depictions, they aim to arrange visual image elements so as to activate viewers' reasoning and inferencing powers, which are to work out the relevance of the image in the verbal and argumentative context.

Owing to the seamlessness of the graphic surfaces as discussed above, it was not always easy to sort images into either the doctored or the un-doctored type. When in doubt, essentially two criteria were applied: First, the archivers' annotation (see Figure 1) helped determine whether a digital artist was involved in production; second, obvious cases of merely retouched images for enhanced beauty and appeal were ignored, but cases of obvious digital manipulation not annotated by the editors were included.

The distribution of the image types in the sampled corpus (see Figure 2) shows that with 41%, the doctored image is in the lead, testifying to the fact that digitally manipulated imagery is fast becoming the technological standard. Classic photography, however is also surprisingly well represented in the corpus with 33%. This demonstrates the stronghold that the indexical, 'documentary' photo still has as a means of depicting fragments of 'reality' in advertising. Illustration, which may be manual or computer-generated, scores 21%, which is perhaps an unexpectedly low proportion considering that many target groups or product types would favor it due to its evocative qualities. The 5% of ads that do not use images at all but



How to use your Archive:
Guide to symbols: ♕: Client ⌂: Advertising Agency 👁: Creative Director ✎: Art Director ✎: Copywriter 📷: Photographer ⚒: Modelmaker ✎: Illustrator Ⓐ: Typographer ✨: Digital Artist ▭: Production Company ▦: Director 🏛: University/School

Fig. 1: Annotation scheme used by *Lürzer's Archive*, containing the 'magic wand'-icon for digital artist.

employ type only prove the observation that current commercial persuasion thrives on very short copy that links with striking imagery, often producing intentionally 'en-riddled' communiqués (cf. Stöckl 2017).

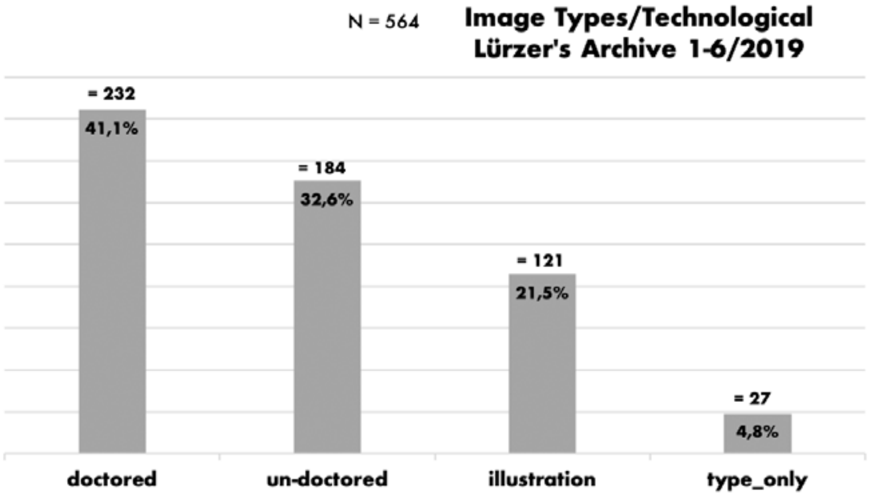


Fig. 2: Breakdown of material-technological image types in the corpus.

3.2 Annotation and Approach

The present study is based on the collection of doctored images as described above, that is research was conducted on a total of 232 advertisements representing 24 product categories. These have all been captured from the archive and been annotated according to the scheme shown in Table 1. Beyond brand and source, the code-system for identifying the ads – e.g., *volkswagen_automobile_argentina_la_2-19_19* – records the product category (*automobile*) and the country of production and distribution (*Argentina*). These criteria may prove useful for quantifying digital design operations in relation to product- and region-specific advertising cultures, a direction not pursued here.

Each advertisement was carefully annotated following the scheme. As the annotation does not capture perceptible formal features but zooms in on content-based and discourse-oriented criteria, it naturally involves individual interpretations by the annotator. I have painstakingly strived for consistency based on establishing analogue cases, and worked over the samples in the corpus multiple times. While I generally trust in the constancy of my classification throughout the corpus, I did have second thoughts about potential uses of multiple annotation, which could have more adequately addressed possible combinations of design operations, rhetorical potentials, and relational propositions. Generally, I found the dimensions and criteria of the annotation easily applicable throughout the corpus and they may, however, have to be adapted for a more large-scale effort.

The present chapter excludes such annotation criteria, such as *text length*, number of *cohesive ties* between text and image elements, and *complexity and semantic basis of ties*, all of which are suited to determining the relative status of text and image, and the exact nature of inter-semiotic coherence. These questions cannot be addressed here for reasons of space. Essentially, three steps have been taken to answer the three research questions outlined in Section 1:

1. Qualitatively, the method seeks to ascertain which *design operations* occur in the corpus and how these can be organized in a taxonomy. Quantitatively, the enquiry draws up a distribution of the design operations in the corpus, which could be indicative of their functional significance in advertising ranging from central to peripheral. Generally, the question answered here concerns the types of manipulations in visual structure.
2. The study further asks which *rhetorical potentials* a design operation has, that is which cognitive-semantic effect derives from the digital manipulation of the graphic material. Such effects occupy an intermediate position between understanding the image and understanding its role in the persuasive argument. Does a given design operation preferably promote a certain rhetorical potential? This question regards the general function a manipulation of visual structure has for how an image is understood in relation to its accompanying copy.
3. As digital design operations ultimately facilitate argument construction, the study asks which conventional *forms of argumentation* design operations feed into. *Relational propositions* are also established for each data set as these build

the basis for multimodal argument construction and shape its specific nature. The question posed here ultimately concerns the kind of logical-semantic relation that image and text enter into for persuasive purposes.

Before cumulative results are presented and interpreted in the light of multimodal commercial argumentation, let us now look at an example in order to capture the descriptive power of the annotation and gauge the assumed interrelations between the elements posited.

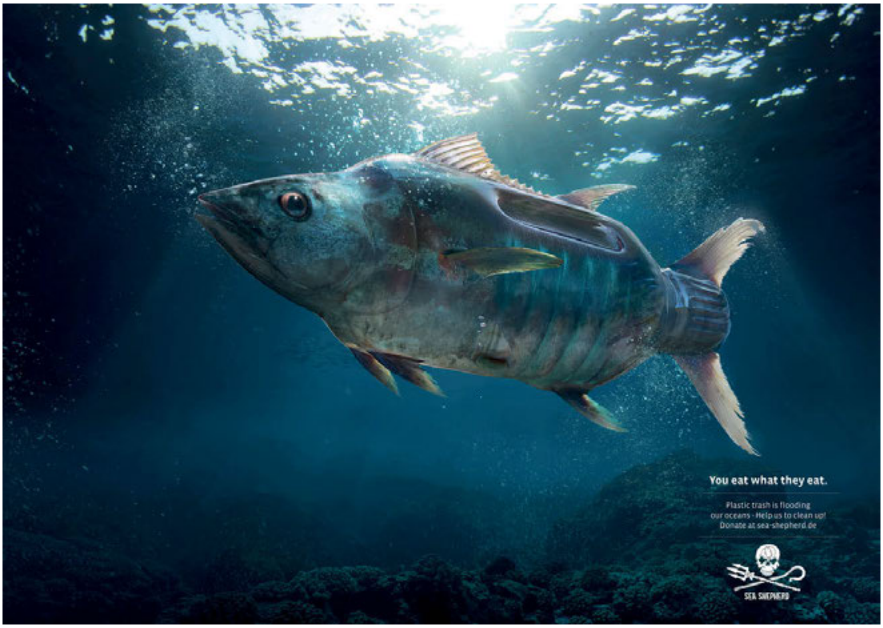

You eat what they eat.
Plastic trash is flooding
our oceans · Help us to clean up!
Donate at sea-shepherd.de
SEA SHEPHERD

Fig. 3: *Sea Shepherd* — Ogilvy, Frankfurt/Main (sea shepherd conservation society_social_germany_la_1-19_102; Lürzer's Archive 1/2019: 102; 1.1907).

The image of the social/environmental ad shown in Figure 3 displays the design operation 'combine', as it intermixes two image elements (fish/plastic bottle) to form a hybrid but unified entity. I labelled the specific sub-type (group) of design operation instantiated here 'morph' because in contrast to 'texture' or 'illude' (see Table 1), the essence of the digital manipulation is to produce an unequivocally recognizable, seamless sign neither based on a merger of surface textures nor on a deliberate optical illusion. Such artful fusion of visual image-elements offers a specific potential for argumentation by suggesting some logical-semantic con-

Tab. 1: Code book for annotation of the corpus. VIE = visual image element.

Annotation Criteria and Codes	Explanations
A — design operation (class)	basic manipulations/configurations of visual structure
1 — *combine*	intermix two or more VIEs to create a hybrid/unified entity/sign
2 — *add/substitute*	modify image by manipulating or exchanging individual VIEs
3 — *juxtapose*	place side by side VIEs of different nature for composite image
4 — *configure type*	fashion type to create VIEs or pictorial qualities central to ad
5 — *model*	3-D-generate an image presenting a scenario or semantic frame
B — design operation (group)	specific manipulations/configurations of visual structure
1 — *combine* >> *a) morph*	fuse two VIEs into a single gestalt by morphing them
1 — *combine* >> *b) texture*	superimpose texture of one VIE over another
1 — *combine* >> *c) illude*	merge VIEs so as to create an optical illusion
2 — *add/substitute* >> *a) corrupt*	distort VIEs into another one or to unusual/divergent shapes
2 — *add/substitute* >> *b) repeat*	reduplicate VIEs to create multiple exemplars
2 — *add/substitute* >> *c) add*	insert VIEs to substitute others or extend them
3 — *juxtapose* >> *a) collage*	montage different image-making materials to produce an image
3 — *juxtapose* >> *b) join*	link two different image types in one digital piece of visual art
3 — *juxtapose* >> *c) relate*	connect two different kinds of pictorial logic (e.g., photo + chart)
4 — *configure type* >> *a) pictorialize*	arrange type so as to compose VIEs or an entire image
4 — *configure type* >> *b) overlay*	superimpose writing over certain VIEs
4 — *configure type* >> *c) associate*	design type so it acquires evocative associative qualities
5 — *model* >> *a) fiction*	generate/arrange 3-D-modelled VIEs into fictitious scenario
5 — *model* >> *b) real*	generate/arrange 3-D-modelled VIEs into seemingly real scenario
C — rhetorical potential	function of manipulation for facilitating multimodal argumentation
1 — *highlight*	make a VIE perceptually salient that is crucial in the argument
2 — *connect*	bring two semantic frames into logical/associative contact
3 — *demonstrate*	illustrate/explain a product's/service's workings
4 — *intensify*	reinforce the intention/impact of a communicative act
5 — *contrast*	compare two entities in terms of similarities/differences
D — argument type	conventionalized rhetorical form of the argument (see Section 4.3)
1 — *example*	image provides illustrative/evidential example of claim in text
2 — *contiguity*	text and image are related through associative/causal logic
3 — *comparison*	text and image show likenesses or contrasts
E — relational proposition	rhetorical structure in the text-image relation (see Section 4.4)
1 — *elaboration*	text specifies/elaborates image or vice versa in various ways
2 — *evidence*	image provides visual evidence for textual claim
3 — *justification*	image justifies the appropriateness/acceptability of speech act
4 — *motivation*	image motivates recipient to comply with directive speech act

nection between the combined signs in terms of contiguity, similarity/analogy, or comparison. This idea of logical or associative connections established by the merger of image elements is labelled 'connect' (see Table 1). In our case, against the backdrop of our world knowledge, the hybridity of the visual gestalt *fish cum bottle* must be read as fish eating plastic and turning into plastic. Exactly this visual idea is taken up and developed by the claim *You eat what they eat*, echoing the saying *You are what you eat* by allusion. When we now attend to multimodal argument construction, it seems obvious that the morphed image serves as an 'example' (see Table 1) of the general verbal claim, both in terms of the fish representing *oceans* and the bottle being an item of *plastic trash*. Finally, text (*Plastic trash is flooding our oceans*) and image can be seen as setting up the relational proposition of 'evidence' (see Table 1), as the image provides visual proof of an inexplicit claim implied by *You eat what they eat*.

The example demonstrates that the method of corpus annotation employed here dissects a holistic product of multimodal communication into assumed units in a process of multimodal argumentation. Fixing *design operation*, *rhetorical potential*, *form of argument*, and *relational proposition* in a larger number of samples clearly allows patterns of multimodal arguments to be discerned.

4 Results and Discussion

What follows is essentially a breakdown of the most prominent design operations (classes and groups) and their rhetorical potentials as well as a survey of common argument types and relational propositions (see Table 2). Relative frequencies are understood here as indications of prototypical and functionally effective image design and multimodal argumentation strategies in advertising. Where possible, I seek correlations between design operations and their assumed rhetorical potentials; these are frequent cooccurrences of such categories. Overall, I attempt an interpretation of the quantitative data along the lines of potential usefulness to the genre and its argumentative tasks.

4.1 Design Operations

With almost 34.9%, the modification and exchange of VIEs (*add/substitute*) is the most widely applied design operation. While reduplicating (*repeat*) and inserting (*add*) VIEs are traditional photoshop operations, distorting (*corrupt*) the shapes of VIEs seems technologically more demanding and, therefore, not surprisingly,

Tab. 2: Breakdown of design operations, rhetorical potentials, argument types and relational propositions in the corpus. AbsF = absolute frequencies; RelF = relative frequencies.

A — Design Operations (class)	AbsF	RelF	B — Design Operations (group)	AbsF	RelF
1 — *combine*	41	17.7%	1a *morph*	23	56.1%
			1b *texture*	5	12.2%
			1c *illude*	13	31.7%
2 — *add/substitute*	81	34.9%	2a *corrupt*	34	42.0%
			2b *repeat*	15	18.5%
			2c *add*	32	40.0%
3 — *juxtapose*	41	17.7%	3a *collage*	5	12.2%
			3b *join*	30	73.2%
			3c *relate*	6	14.6%
4 — *model*	54	23.3%	4a *fiction*	28	51.9%
			4b *real*	26	48.1%
5 — *configure*	15	6.5%	5a *pictorialize*	6	40.0%
			5b *overlay*	8	53.3%
			5c *associate*	1	6.7%

C — Rhetorical Potential	AbsF	RelF
connect	59	25.4%
contrast	21	9.1%
demonstrate	61	26.3%
highlight	71	30.6%
intensify	20	8.6%

D — Argument Type	AbsF	RelF
comparison	53	22.8%
contiguity	94	40.5%
example	85	36.6%

E — Relational Proposition	AbsF	RelF
elaboration	102	44.0%
evidence	80	34.5%
justification	5	2.2%
motivation	45	19.4%

is the leading sub-type within the *add/substitute* operation (42.0%). A trend towards complex and creatively challenging design operations is corroborated by the fact that 3D-image generation comes second (23.3%). It is in the modelling of images from scratch in order to engineer semantic frames or scenarios that digital art shows its true potential. Virtually any argumentatively useful image can be created through 3D-models. Interestingly, *fictitious* (51.9%) and *real* (48.1%) visual scenarios are almost evenly distributed in the corpus.

Models generally hold strong visual appeals and enjoy a considerable propositional and argumentative flexibility, while their seamlessness is maximal and their hyper-realness facilitates tremendous ideational freedom. The two design operations *combine* and *juxtapose* are on par (almost 20% each) and constitute the third most frequent type of image manipulation. They share a similar intention in bringing together disparate and semantically incongruous, but ultimately relatable entities. While combining aims to create unified, largely seamless visual entities, *juxtaposing* seeks to make the linking of heterogeneous materials, image types, or logics salient. Within the first class, *morphing* (56.1%) clearly proves to be the most dominant group of design operation, entailing the fusing of two VIEs into a single gestalt or supra-sign (see Figure 3).

One may hypothesize that such morphed images necessitate comparatively high levels of cognitive involvement, as do deliberate optical illusions (*illude*/31.7%). With both, playfulness and creative scope are at their best. Juxtapositions of different image types (*join*/73.2%), e.g., b/w or colour photography, painting, x-ray, etc., prove most significant in the corpus, which may simply be due to the comparative ease with which such operations may be applied. Juxtaposing differing materials (*collage*/12.2%) and logics (*relate*/14.6%) appears to be much harder. While the argumentative usefulness of combining lies in pointing out likeness or logical relatedness, juxtaposing often initiates comparison that leads to the identification of semantic differences.

Finally, while being relatively infrequent in the overall data, configurations of print (*configure*/6.5%) are interesting to look at in detail and demonstrate the not altogether negligible impact typography may have on argumentative text-image relations (Stöckl 2014b). Two groups of configuring operations seem to particularly support multimodal argumentation: First, *overlay* (53.3%) strategically places writing on VIEs to emphasize or modify its propositional content; second, *pictorialize* (40.0%) arranges type so that it transmutes into VIEs or an entire image. The latter is creatively and technologically very demanding and very likely creates high levels of attention and appeal.

4.2 Rhetorical Potentials

When we survey rhetorical potentials, our focus shifts to the function of design operations as manipulations of visual structure for facilitating multimodal argumentation. Altogether five potentials have been distinguished and three of them seem relatively more significant in the corpus than the two remaining ones. With 30.6%, *highlighting* is the most dominant rhetorical potential, which entails making a VIE perceptually salient that is crucial to the argument to be constructed. The design operation that most strongly facilitates highlighting is clearly add/substitute, as here individual VIEs are the object of manipulation and anything added, substituted, repeated or formally corrupted is likely to stand out. So, for instance, a recruitment ad for the fire brigade (atlanta fire rescue foundation_social_usa_la_1-19_108) highlights the diversity of fire extinguishers by repeating this VIE in six different colors. The argument constructed multimodally that staff may/should be as diverse as fires (*Fires are indiscriminate. So are we.*) thrives on the design operation *add/substitute – repeat* and its highlighting function. Other design operations may also promote the rhetorical potential of highlighting: Juxtaposing does so by making the contrast between image materials, types, and logics salient; configuring print may direct perceptual attention to the pictorial qualities of writing.

The two rhetorical potentials *demonstrate* (26.3%) and *connect* (25.4%) are the second most significant in the corpus. The first illustrates or explains the quality or workings of a product and thus seems at the heart of commercial argumentation. The latter brings two semantic frames/elements into logical or associative contact in order to support an argument. While the rhetorical function of demonstration correlates most strongly with the design operation model, connect is often realized by the design operation *combine*. For instance, an ad for a specially tuned high-performance GT car (mercedes_automobile_china_la_4-19_15; see Figure 4) shows the plausible interrelation between model and demonstrate. It uses a 3-D model that provides a scenario of a fantasy creature motherly caring for its off-spring. This digital art demonstrates the claim *the beast is now family friendly*, with reference to the product quality *4 doors* and a visual allusion perhaps to *The Gremlins*. An ad for a beauty clinic (afine estética_services_brazil_la_2-19_79) offering *cryolipolysis*, i.e., *fat-freeze weight loss*, by comparison, instantiates the coupling of *combine* and *connect*. Here, the image superimposes the texture of one VIE (ice) over another (a burger) and thus connects two semantic frames (freezing and eating) to prepare the visual ground for the (counter-)claim that fat-freezing works (*They say cold makes us fat, we proved them wrong*).

Finally, two minor rhetorical potentials of design operations in the corpus turned out to be *contrast* (9.1%) and *intensify* (8.6). They both correlate with the design operation *juxtapose*, which for its link to *contrast* is obvious as comparison requires the placement of two VIEs in close proximity or combination. The connection between *intensify* and *juxtapose* is less straightforward. Mostly, in these cases, the impact of a verbal claim or a directive speech act (recommendation,

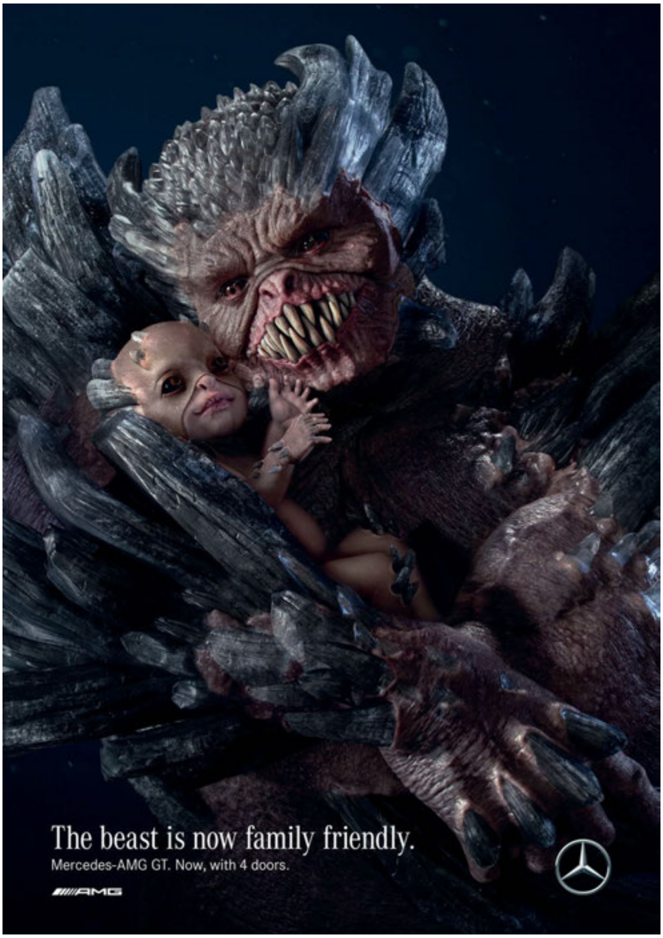

The beast is now family friendly.
Mercedes-AMG GT. Now, with 4 doors.
AMG

Fig. 4: *Mercedes* — BBDO, Shanghai (mercedes_automobile_china_la_4-19_15; Lürzer's Archive 4/2019: 15; 4.1921).

call for action) is reinforced by digital art that supports an implicit comparison and serves as evidence for the claim or motivation for the directive (see relational propositions, Section 4.4). An ad for mortadella (frisa_food_brazil_la_2-19_38) serves as a typical example: the direct-address claim *Frisa is the size of your hunger* is supported by a composite image which juxtaposes product photography (the sausage) and mural-style illustration (the fictitious character eating the meat) to prove the extraordinary size of the sausage and emphasize the direct address.

4.3 Argument Types

It is common in rhetorical theory to distinguish different types or forms of arguments (Lehn 2011; Janich 2013: 131–136). These essentially represent more or less distinct patterns of configuring claims, premises, and conclusions into an argumentative structure. I have extended this view by regarding them as multimodal on principal, i.e., images and their elements will contribute to the overall argument. While their delineation may not be watertight and subdivisions are possible, three broad types of arguments have been found to be central in the corpus: *contiguity* (40.5%), *example* (36.6%), and *comparison* (22.8%). As no correlations between these argument types and design operations have been explored yet, I must assume so far that all three of them are typical of advertising as a genre. This, in turn, means that their distribution is perhaps ad-specific but not necessarily connected to the various uses of digital art.

First, *contiguity-based arguments* employ an associative or essentially causal logic to connect textual and image elements. For example, an animal protection ad (freeland_social_thailand_la_5-19_86) shows a bleeding deer hanging from the outside wall of a posh apartment building and relates this to the claim *the truth behind the wall*. In order to construct the multimodal argument, metonymic relations (body on the outside for antlers on the inside; apartment for people living in it) and causal relations (hunting for trophies kills animals) need to be established. Second, *example-based arguments* use digital art to construe an illustrative, specific evidence for general claims made in the text. So, for instance, a watch ad (everlast_accessories_brazil_la_5-19_13) specifies and illustrates the claim *For those who don't waste time with shallow water* by showing the tentacle of an octopus in the water wearing the watch. Finally, *comparison-based arguments* involve text and VIEs in the construction of all kinds of differences and similarities. An ad promoting *powerful batteries* (nanfu_house_china_la_4-19_70), for instance, employs a realistic 3-D model of two sumo wrestlers in order to metaphorically compare electrical to physical power. In summary, it can safely be said that *contiguity* is largely based on metonymic and causal text-image relations, *examples* presuppose some general-specific relation, and *comparison* is supported by analogical, contrastive, and metaphorical text-image relations.

4.4 Relational Propositions

In an attempt to capture the multimodal workings of commercial arguments, rhetorical structure theory (Mann & Thompson 1988) generally proves helpful because it models discourse as relations between propositional parts of text. If we allow for images as essential propositional elements in a multimodal discourse structure (cf. Bateman 2014: 213–220), we can start to set up relational propositions between linguistically and pictorially realized content. From the catalogue of relations (Mann & Thompson 1986), four featured in the corpus to varying degrees — all of them can be shown to be crucial to the kinds of multimodal arguments to be developed in advertisements.

With 44.0%, the relational proposition elaboration turns out to be the most vital in the corpus. It is at work, for instance, in a health-food restaurant ad (sana sana_food_costa rica_la_1-19_57, see Figure 5) that shows a celery morphed into an x-ray image of bones and uses the text: *A celery has the calcium that your bones need. Sana Sana. Food heals*. Here, the text elaborates the image by establishing an associative or even causal connection between celery/greens (containing calcium) and hard/healthy bones. The centrality of elaboration in the corpus can be explained in two ways: First, apparently, the propositions relayed through digital art require explanation by verbal contextualization and second, elaboration is a very versatile and consequently underspecified relational proposition that contains five subtypes under the rubric of a general-specific relation (Mann & Thompson 1986: 63–64). The directionality appears to be mainly text elaborating an image, but the reverse also occurs in the corpus when there is little text and a potent image.

The relational proposition *evidence* is the second most frequent in the corpus (34.5%). Here, an image provides visual evidence for a textual claim as in a banking ad (banque_richelieu_banking_france_la_2-19_23) that visualizes the claim *Banque Richelieu. The spirit of adventure since 1624* by showing a high-speed train in glass tubes racing through the sylvan landscape of a baroque painting. Fictitious or real 3-D-modeled scenarios would seem to be particularly useful in providing striking visual evidence, but other design operations can be equally useful in realizing this strongly ad-typical relational proposition, as the evidence function generally relies on the indexicality and iconicity of images. Regarding argument types, evidence could plausibly have a strong correlation with examples. Whether the evidence provided in the image is realistic and trustworthy is clearly less relevant than the creativity, seamlessness, and freshness of the digital art that goes into this relational proposition. Reversals in the directionality of text and image seem counterintuitive for the claim-evidence relation.

Represented with 19.4%, *motivation* covers multimodal arguments where an image is to motivate a recipient to comply with a directive speech act, which may be either a recommendation (*Have a break, have a Kitkat*) or a warning (*Don't text and drive*). A Kitkat-ad (kitkat_food_greece_la_2-19_33), for instance, motivates the famous imperative slogan by showing a diving board shaped like a Kitkat bar in an outdoor swimming pool. Here, the visual motivation depicts an ideal or desirable situation that the product fits in or facilitates. On the other hand,

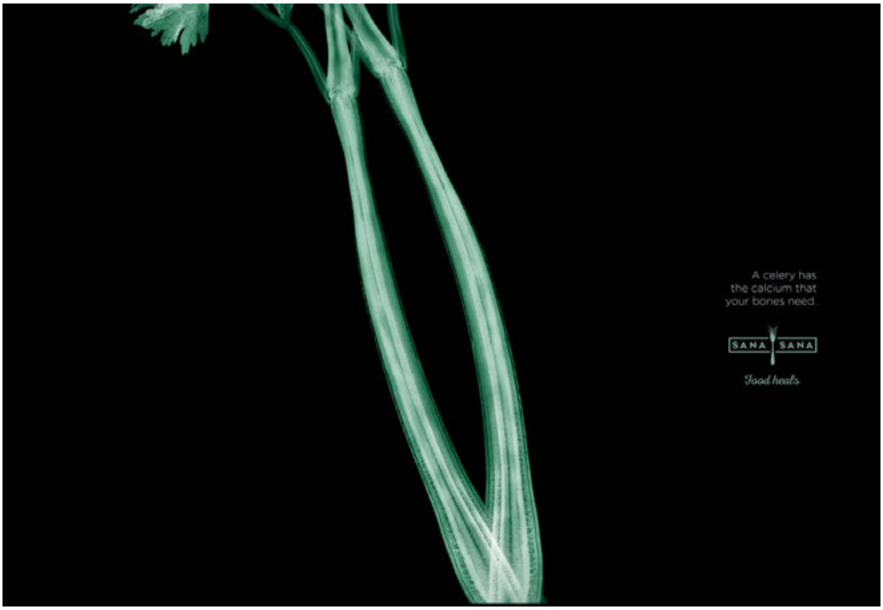


Fig. 5: *Sana Sana Restaurant* — Gitanos, San Jose (sana sana_food_costa rica_la_1-19_57; Lürzer's Archive 1/2019: 57; 1.1910).

warnings can be motivated by showing contexts of danger or dilemma, which the advertised product/service promises to eliminate or alleviate. For instance, a banking ad (jordan_commercial_bank_banking_jordan_la_2-19_22) illustrates the warning *Don't blow your money away on rent* by a 3-D-modeled image of a house literally dissolving into bits of rubble and dust blown away by a storm. The samples demonstrate that motivation is at home both in commercial and social advertising; its essence lies in giving the recipient an argumentative incentive for future ad-compliant action.

Finally, with only 2.2%, *justification* appears to be a negligible relational proposition, where an image serves to legitimize the appropriateness of the very speech act that is addressed to the recipient as if immediate interaction was possible in a mass-mediated context. So, for instance, a pasta-sauce ad (rao's homemade_food_usa_la_2-19_36) asks the insistent question *What homemade sorcery is this???* and justifies it by an elaborate *juxtaposition* of photography and drawing (cf. the pun on *sorcery/sauce*, which also refers to the image-type). Speech acts that require especial justification would seem to be all kinds of impositions or apologies for inappropriate address and content.

5 Summary and Conclusion

The present study has yielded a number of tenable results that would merit further investigation. First of all, a workable and empirically verified taxonomy of design operations was established as a system of classes and groups that designate manipulations of visual structure, style, and content. The prominence of 3-D-modeling and combining of VIEs in the corpus points to the specific creative scope that print-CGI offers and indicates that, indeed, the new medium is increasing the flexibility and freedom of image-making. Whether truly new ideas for commercial argumentation are emerging owing to the rhetorical potential of print-CGI would need further enquiry. However, the data suggests that it is especially fictitious models, the morphing of visual signs and optical illusions whose seamlessness creates new visual and argumentative options. The most pervasive rhetorical potentials, i.e., the functions the design operations fulfill in facilitating multimodal argumentation, appear to be the highlighting of VIE that are crucial for argument construction and visually explaining the workings of a product/service. These and other rhetorical potentials for creating a semantic-cognitive effect of the visual in multimodal argumentation translate into three basic types of arguments: metonymically and causally based contiguity, examples that specify a general claim, and comparisons that rely on analogy, contrast or metaphor. Finally, an examination of the multimodal relational propositions at work in commercial argumentation suggests elaboration, visual evidence, and motivation as the most potent discourse patterns, whose directionality or text-image centricity varies.

The method employed may both be appreciated for its powers and scrutinized for its limitations. It deliberately avoids working with established rhetorical figures because these are overcomplex, fuzzy, and hard to handle in analysis. Instead, digitally performed manipulations of graphic material are observed and classified in order to relate them to their potential function in types of arguments and types of multimodal discourse relations. This approach circles in on the nature of multimodal argumentation from a number of partly overlapping but essentially complementary angles. Its analytical categories aim at making prominent multimodal, discursive, and argumentative patterns visible in the data. However, it is a natural drawback of this analytical method that it backgrounds the complex and multifarious ways in which contextually relevant interpretations of multimodal advertisements are construed. Forceville (2020: 80–95) rightly points out that multimodal communication can only be adequately explained as an act of creating relevance through multiple pragmatic procedures and cognitive operations, such as, for instance, explicatures/implicatures and metaphor/metonymy.

The annotated corpus offers scope for further related explorations. First, significant correlations between the analytical categories, e.g., between design operations and rhetorical potentials or between arguments types and relational propositions, could be examined. This would lead to a more holistic interpretation of the data and help attain a less atomistic view of multimodal argumentation that could demonstrate the co-dependency of the parameters involved. Most importantly, using such annotation criteria as the number of ties between VIEs and linguistic expressions plus the nature of their inter-semiotic sense relations, we could describe the relative status of text and image and their logical-semantic relations in detail. Also, typical patterns of argumentation could be charted in relation to product type or advertising cultures. Finally, in order to improve the statistical robustness of the study and its results, the appropriateness of its method could be enhanced by multiple independent annotators and by establishing their inter-annotator agreement as well as by all kinds of validity-testing (cf. the criteria and tests suggested by Bateman & Hiippala (2020: 3, 8–16)).

Bibliography

Bateman, John. 2014. *Text and Image: A Critical Introduction to the Visual/Verbal Divide*. London: Routledge.

Bateman, John A. 2018. Position Paper on Argument and Multimodality. Untangling the Connections. *International Review of Pragmatics* 10. 294–308. https://doi.org/10.1163/18773109-01002008.

Bateman, John A. & T. Hiippala. 2020. Statistics for Multimodality: Why, When, How – An Invitation. SocArXiv https://doi.org/10.31235/osf.io/7j3np.

Bühler, P., P. Schlaich & D. Sinner. 2017. *Digitales Bild. Bildgestaltung – Bildbearbeitung – Bildtechnik*. Berlin: Springer.

Durand, Jacques. 1987. Rhetorical Figures in the Advertising Image. In J. Umiker-Sebeok (ed.), *Marketing and Semiotics: New Directions in the Study of Signs for Sale*, 295–318. Berlin, New York and Amsterdam: De Gruyter Mouton.

van Eemeren, F.H. & B. Garssen (eds.). 2012. *Topical Themes in Argumentation Theory: Twenty Exploratory Studies*. Dordrecht: Springer.

Forceville, Charles. 2020. *Visual and Multimodal Communication. Applying the Relevance Principle*. Oxford: Oxford University Press.

Forceville, Charles J. 2014. Relevance Theory as Model for Analyzing Visual and Multimodal Communication. In D. Machin (ed.), *Visual communication*, 51–70. Berlin: De Gruyter Mouton.

Gaede, W. 1981. *Vom Wort zum Bild. Kreativ-Methoden der Visualisierung*. München: Langen Müller/Herbig.

Janich, Nina. 2013. *Werbesprache. Ein Arbeitsbuch. 6th edition*. Tübingen: Narr.

Kjeldsen, Jens E. 2012. Pictorial Argumentation in Advertising: Visual Tropes and Figures as a Way of Creating Visual Argumentation. In F. H. van Eemeren & B. Garssen (eds.), *Topical* *themes in argumentation theory: twenty exploratory studies* (Argumentation Library 22), 239–255. Berlin: Springer.

Lehn, Isabelle. 2011. *Rhetorik der Werbung: Grundzüge einer rhetorischen Werbetheorie.* Konstanz: UVK.

Lürzer's International Archive. 2013/2014. *200 Best Digital Artists Worldwide.* Vienna, etc.

Lürzer's International Archive. 2019/2020. *200 Best Digital Artists Worldwide.* Vienna, etc.

Mann, William C. & S. A. Thompson. 1986. Relational Propositions in Discourse. *Discourse Processes* 9(1). 57–90.

Mann, William C. & S. A. Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text* 8(3). 243–281.

McQuarrie, Edward F. & D. G. Mick. 1996. Figures of Rhetoric in Advertising Language. *Journal of Consumer Research* 22(4). 424–438.

McQuarrie, E.F. & B. Philipps. 2005. Indirect Persuasion in Advertising: How Consumers Process Metaphors Presented in Pictures and Words. *Journal of Advertising* 34. 7–20.

van Mulken, M., R. van Enschot-van Dijk & H. Hoeken. 2005. Puns, Relevance and Appreciation in Advertisments. *Journal of Pragmatics* 37. 707–721.

Müller, Marion G. & S. Geise. 2015. *Grundlagen der visuellen Kommunikation. Theorieansätze und Methoden.* Konstanz: UVK.

Phillips, Barbara J. & E. F. McQuarrie. 2004. Beyond Visual Metaphor: A New Typology of Visual Rhetoric in Advertising. *Marketing Theory* 4(1/2). 113–136.

Roque, G. 2012. Visual Argumentation. A Further Reappraisal. In F. H. van Eemeren & B. Garssen (eds.), *Topical Themes in Argumentation Theory*, 273–288. Dordrecht: Springer.

Scholz, O. 1998. Was heißt es, ein Bild zu verstehen? In K. Sachs-Hombach & K. Rehkämper (eds.), *Bild – Bildwahrnehmung – Bildverarbeitung. Interdisziplinäre Beiträge zur Bildwissenschaft*, 105–117. Wiesbaden: UVK.

Stöckl, Hartmut. 2004. *Die Sprache im Bild — Das Bild in der Sprache: Zur Verknüpfung von Sprache und Bild im massenmedialen Text. Konzepte – Theorien – Analysemethoden.* Berlin: De Gruyter Mouton.

Stöckl, Hartmut. 2009. Beyond Depicting. Language-Image-Links in the Service of Advertising. *AAA – Arbeiten aus Anglistik und Amerikanistik* 34. 3–28.

Stöckl, Hartmut. 2014a. Rhetorische Bildanalyse. In N. Bildphilosophie (ed.), *Bild und Methode: Theoretische Hintergründe und Methodische Verfahren der Bildwissenschaft*, 377–390. Köln: Halem.

Stöckl, Hartmut. 2014b. Typography. In S. Norris & C. D. Maier (eds.), *Interactions, Images and Texts. A Reader in Multimodality*, 281–296. Berlin and Boston: De Gruyter Mouton.

Stöckl, Hartmut. 2017. The Multimodal Enigmatic Advertisement – 'En-riddling' as a Rhetorical Strategy in Commercial Persuasion. In P. Handler, K. Kaindl & H. Wochele (eds.), *Ceci n'est pas une festschrift. Texte zur Angewandten und Romanistischen Sprachwissenschaft*, 69–81. Berlin: Logos.

Tseronis, Assimakis & C. Forceville (eds.). 2017. *Multimodal Argumentation and Rhetoric in Media Genres.* Amsterdam: John Benjamins.

Tseronis, Assimakis & C. Pollaroli. 2018. Pragmatic Insights for Multimodal Argumentation. *Special Issue of International Review of Pragmatics* 10.

Jiaping Kang and Zhanhao Jiang

Multimodal Discourse Analysis Based on the GeM Model

A Case Study of Environment Protection Posters

Abstract: Multimodal research is naturally interdisciplinary, allowing research studied from different disciplines and theoretical perspectives. The main challenge is the lack of empirical research with large corpora, especially when applied to visuo-verbal discourse. This chapter takes the GeM (Genre and Multimodality) model as the theoretical framework to analyze environment protection posters from China and the USA. The data set consists of ten posters, with five posters in each sub-corpus, from governmental agencies and organizations. The GeM model adopts XML (eXtensible Markup Language) to annotate the rhetorical structure of the posters, and uses GeM-Tools to conduct the research. The research analyzes the semiotic resources (language, images, and layout) participating in meaning generation to investigate how modes construct meaning together, and to clarify the typical features and relations between language and images of environment protection posters as multimodal discourses. Through an empirical research, the chapter undertakes a contrastive analysis of characteristics in semiotic resources of Chinese and American environment protection posters, conducts statistical tests to examine the significance of the differences, and explores the reasons for differences from a cross-cultural perspective.

Keywords: multimodal discourse analysis, GeM model, RST, environment protection posters

1 Introduction

Since the 1970s, multimodal discourse analysis has been conducted from several perspectives, including social semiotics, cognitive linguistics, interaction analysis, conversational analysis and pragmatics. However, there are still many limitations in the theoretical construction and research methods of multimodal discourse analysis. The main challenge is a lack of empirical research based on sufficiently large corpora (Bateman et al. 2017), especially when applied to visual-verbal discourse.

The Genre and Multimodality (GeM) model (Bateman 2008), as the first framework within linguistics for a multi-level description and analysis of multimodal visually-organized static documents/artifacts, provides an effective tool for anno-

https://doi.org/10.1515/9783110725001-009

tating print media discourses. This framework has mainly been used in the study of XML-based (eXtensible Markup Language) multimodal corpora built for spoken language. Only few scholars have investigated corpora of graphic artefacts. Three cases of XML-based multimodal corpora using the GeM-scheme are mentioned here. Thomas (2009) conducted a study of localization on the basis of a corpus of 24 product packaging from the UK and Taiwan. Hiippala (2014) studied the interaction of multiple semiotic resources of annotated 58 double-paged tourist brochures of Helsinki published between 1967 and 2008. Zhang (2018) used the GeM framework to compare a total of 60 public health posters of New York and Hong Kong.

The research reported here will take the GeM model as theoretical framework to analyze environment protection posters from China and America. This model adopts XML to annotate the rhetorical structure of posters and uses the GeM-Tools (Hiippala 2015b) to conduct the research. The research analyzes the semiotic resources (language, images, and layout) participating in meaning generation to investigate how modes construct meaning together and to clarify the typical features of the genre of environment protection posters as multimodal discourse. Through this empirical research, the chapter undertakes a contrastive analysis of characteristics in semiotic resources of Chinese and American environment protection posters, conducts statistical tests to examine the significance of the differences, and explores possible reasons for differences from a cross-cultural perspective.

2 Theoretical Framework

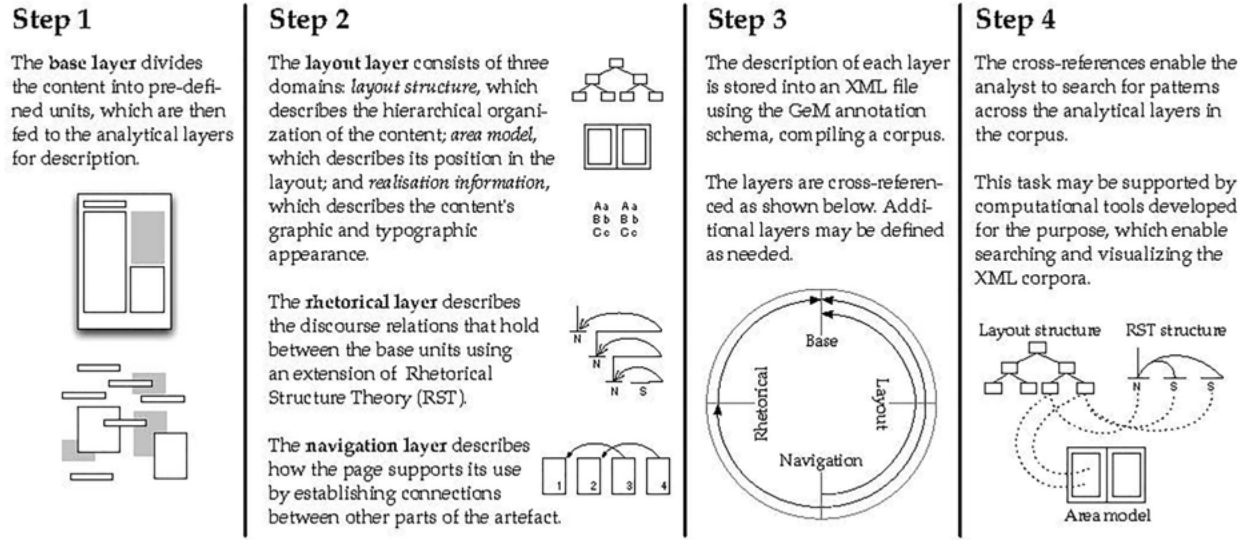
2.1 The Genre and Multimodality (GeM) Model

The GeM model treats multimodal documents as multi-layered artifacts. It provides an annotation scheme with four main analytical layers: the base layer, the layout layer, the rhetorical layer, and the navigation layer. **The base layer** carries the forms that appear regardless of their modes of expression and acts as the starting point for analysis. It provides a list of Recognized Base Units (RBUs), which can be referenced in the subsequent analytical layers (Bateman 2008: 111). The base layer explicitly identifies all the base units that are to be treated as atomic for the purposes of analysis and labelling. Sentences, photos, captions, list items and forms in other semiotic resources are all usually identified as base units. **The layout layer** describes the hierarchical visual clustering of the identified base units, their layout positioning, and their typographic or graphic features. The layout layer consists of three subcomponents: the layout structure itself, which describes the hierarchical visual structure holding over layout units; the area model, which describes the relative placement of layout units; and realization information, which describes the typographic and graphic characteristics of layout units. **The rhetorical layer** describes the rhetorical relations between content elements on a page by using a revised version of Rhetorical Structure Theory (Taboada & Mann 2006) extended to cover visual elements in order to analyze text-image relationships. **The navigation layer** describes navigational structures by defining pointers, entries, and indices which facilitate the use of the document. Navigation elements can be realized by both verbal and visual elements, such as page numbers or arrows. Overall, it is worth mentioning that the analytical layers are not limited to those mentioned above; analysts can add more if needed. Hiippala (2017) provides an overview of annotation layers and visualizes the analytical process in applying the GeM model to support empirical research on page-based documents as shown in Figure 1.

2.2 Rhetorical Structure Theory

2.2.1 A Brief Introduction

Rhetorical Structure Theory (hereafter RST) is a theory of text organization, which was originally developed by Mann & Thompson (1987, 1988). RST aims to explain the coherence of text by describing the relations holding between its components (text spans); minimal spans are called units (Mann & Matthiessen 1991: 233). RST defines a basic set of 24 rhetorical relations (21 nucleus-satellite relations and 3 multinuclear relations) at the beginning, which was expanded later as it is in principle open-ended. Based on the relations listed on the RST website developed by Taboada and Mann (http://www.sfu.ca/rst/), more relations from other scholars (Matthiessen 2002, 2014; Matthiessen & Teruya 2015; Zhang 2018) are added to the list. 25 nucleus-satellite relations include antithesis, background, circumstance, concession, condition, elaboration, enablement, evaluation, evidence, interpretation, justify, motivation, otherwise, purpose, restatement, solutionhood, summary, cause, result, preparation, means, manner, unconditional, unless, and projection. Eight multinuclear relations include addition, conjunction, contrast, joint, list, sequence, multinuclear restatement, and multinuclear projection. The nucleus-satellite scheme is referred to as asymmetric and the multinuclear as symmetric.



Step 1

The **base layer** divides the content into pre-defined units, which are then fed to the analytical layers for description.

Step 2

The **layout layer** consists of three domains: *layout structure*, which describes the hierarchical organization of the content; *area model*, which describes its position in the layout; and *realisation information*, which describes the content's graphic and typographic appearance.

The **rhetorical layer** describes the discourse relations that hold between the base units using an extension of Rhetorical Structure Theory (RST).

The **navigation layer** describes how the page supports its use by establishing connections between other parts of the artefact.

Step 3

The description of each layer is stored into an XML file using the GeM annotation schema, compiling a corpus.

The layers are cross-referenced as shown below. Additional layers may be defined as needed.

Step 4

The cross-references enable the analyst to search for patterns across the analytical layers in the corpus.

This task may be supported by computational tools developed for the purpose, which enable searching and visualizing the XML corpora.

Fig. 1: Methodological steps in applying the GeM model (taken from Hiippala 2017: 278).

2.2.2 Multimodal RST

RST is not sufficient to describe multimodal documents, so relations have been added to deal with both verbal and graphical elements. Bateman discusses the move to multimodal RST and also highlights several problems related to GeM RST. The first problem lies in the spatial relations of segments in multimodal documents, in contrast to the sequentiality of text segments conventional RST builds on. In order to solve this problem, Bateman (2008: 158) proposes that RST relations are restricted to pairs of document parts adjacent to each other in any direction (cf. left, right, upper, and lower neighboring segments). A second problem relates to the nuclearity assignments between verbal and visual segments in multimodal documents. In order to avoid arbitrary assignments of nuclearity in image-text relations, Multinuclear restatement relations and multinuclear Projection relations are employed (Bateman 2008: 159). A third challenge is related to the fact that one segment may serve more than one purpose in a single document and therefore stands in an RST relation to more than one document-part at a time. In RST, spans may not be reused, but an image can be reused in more than one role in a rhetorical organization. In order to solve the problem, we maintain "the tree-like hierarchical organization of the RST analysis by specifying and refining just what it means for elements to be the 'same' or not" (Bateman 2008: 159). Finally, one further problem concerns the adopted minimal unit of analysis in multimodal rhetorical analysis. In order to account for subnuclear elaboration relations that would have been expressed by clauses of "being" or "possession", the GeM model adds five intra-clausal relations based on Halliday & Matthiessen (2013: 210–248) and Bateman (2008: 162) to extend its analytical reach, that is, identification, class-ascription, property-ascription, possession, and location.

RST has been widely used for a variety of research fields, "ranging from linguistic text interpretation to computational applications in language production and automatic analysis" (Bateman & Delin 2006). In the present study, RST is used to make the rhetorical relations of different semiotic resources explicit, and analyze how language and images work individually and work together to generate meaning in Chinese and American environment protection posters.

3 Building the Multimodal Corpus

3.1 Data Collection

The corpus analyzed here consists of ten environment protection posters, shortened as CEPP, as shown in Figure 2, which were collected from March 2019 to August 2019. The ten environment protection posters are sampled in equal numbers from China and America, CEPP-CN and CEPP-US respectively, with five posters in each sub-corpus. The data is collected from governmental agencies and organizations in China and America, including the social media channels (e.g., Facebook, Twitter). Table 1 lists the data sources and the number of environment protection posters collected from those sources. In order to make the corpus more representative, we adopted both systematic sampling and random sampling selection methods (Bateman et al. 2017). First, the corpus was systematically organized by topics; the posters on the same/similar topic are marked with the same Arabic numerals, as shown in Figure 2. Second, posters sampled for the same/similar environment topic were randomly selected. According to the Ministry of Ecology and Environment (MEE) [1] and the Environmental Protection Agency (EPA)[2], air, water, green living, waste, and biodiversity are among the top ten environmental issues. The CEPP corpus in the present research is collected by covering the above topics, including Green commuting, Energy saving, Water resource management, Recycling and Biodiversity.

3.2 Multimodal Corpus Annotation

3.2.1 GeM Annotation Scheme

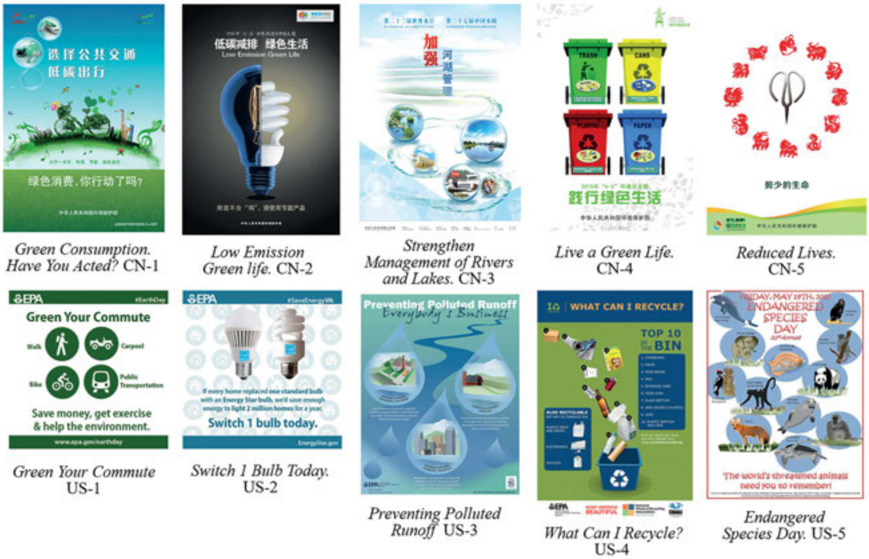
The GeM annotation scheme uses the eXtensible Markup Language (XML) for organizing the annotation as described in Bateman (2008) and Hiippala (2015a). As a markup language, XML uses tags to annotate and defines a set of rules for encoding documents. The tags of XML are not pre-given, which allows researchers

[1] Ministry of Ecology and Environment (MEE) implemented the central committee's guidelines, policies and arrangements on ecological environmental protection, and upheld and strengthened the party's centralized and unified leadership over ecological environmental protection in China. https://www.mee.gov.cn/.

[2] Environmental Protection Agency (EPA) is an independent administrative agency which is responsible for protecting the natural environment and human health from environmental hazards in America. https://www.epa.gov/.

Tab. 1: Data sources and numbers of the environment protection posters.

CEPP	Source	Number
CEPP-CN(5)	Ministry of Ecology and Environment of the PRC(MEE)	1
	Center for Environmental Education and Communications of MEE	1
	Ministry of Water Resources of the PRC	1
CEPP-US(5)	United States Environmental Protection Agency (EPA)	3
	Energy Star	1
	EPA & National Waste & Recycling Association & Solid Waste Association of North America (SWANA)	1



Green Consumption. Have You Acted? CN-1

Low Emission Green life. CN-2

Strengthen Management of Rivers and Lakes. CN-3

Live a Green Life. CN-4

Reduced Lives. CN-5

Green Your Commute US-1

Switch 1 Bulb Today. US-2

Preventing Polluted Runoff US-3

What Can I Recycle? US-4

Endangered Species Day. US-5

Fig. 2: Collage of ten environment protection posters.

to define their own. In the multi-layered annotation process, different sets of XML tags, defined according to schemes, are employed to annotate different layers. According to Henschel's (2003) GeM annotation manual, the base units, layout units and basic RST units are marked by <unit>, <layout-unit> and <segment> tags, respectively. Each individual analytical unit is given a unique identifier. For instance, the prefix '*u*-' in u-01.01 indicates that it is an identifier of the base layer. As the annotation for each analytical layer is stored in its own file, file names should be distinguished from each other among multiple files. The labeling scheme adopted in this chapter is presented in Table 2.

Tab. 2: Tags, identifiers, and file names of each layer.

Analytical layer	Tag	Identifier	File name
Base layer	`<unit>`	`u-01.01, u-01.02...`	US-base-1.xml
Layout layer	`<layout-unit>`	`layout-unit id`	US-lay-1.xml
	`<segment>`	`s-01.01, s-01.02...`	

RST layer	`<span>`	`span-01.01,`
		`span-01.02...`
			US-rst-1.xml

	`<multi-span>`	`span-01.01,`
		`span-01.02...`	

Navigation layer	`<pointer>`	`p-01.01, p-01.02...`	US-nav-1.xml
	`<entry>`	`e-01.01, e-01.02...`	

Since navigation devices are seldom used in the present study, the following description will focus on the annotation of three layers: the base layer, layout layer, and rhetorical layer. In order to specify each element, different XML attributes are employed. The attribute `id` carries a unique identifying symbol used to distinguish each base unit, layout unit and RST unit. The attribute `xref` refers to the base units which belong to a layout unit or an RST unit. In addition, posters as multimodal artefacts consist of both verbal and visual elements. For units consisting of verbal elements, `<unit>`, `<layout-unit>` and `<segment>` tags are used. For photographs and other graphic elements, the attribute `alt` is employed in the annotation so as to describe the content of the visual elements. Here are some examples:

```
<unit id=”u-01.03''> Green Your Commute</unit>
<layout-unit id=”lay-01.03” xref=”u-01.03''> Green Your Commute
</layout-unit>
<segment id=”s-01.01” xref=”u-01.03”/>
<unit id=”u-02.05” alt=”Image: Half a tungsten bulb”/>
<segment id=”s-02.04” xref=”u-02.05” alt=”Image: Half a tungsten
bulb”/>
<layout-unit id=”lay-02.03” xref=”u-02.05” alt=”Image: Half a
tungsten bulb”/>
```

The XML annotation of rhetorical structure is specified by the analysis of rhetorical relations. The three categories of relations mentioned above, that is, nucleus-satellite relations, multinuclear relations, and intra-clausal relations, are marked as `span`, `multi-span`, and `mini-span` respectively. Annotation examples are as follows:

```
<span id=”span-04.01” nucleus=”s-04.01” satellites=”span-04.02”
```

```
relation="solutionhood"/>
<multi-span id="span-04.05" nuclei="span-04.06 span-04.07 span-04.08"
relation="joint"/>
<mini-span id="span-04.06" attribute="s-04.40" attribute="s-04.41"
relation="class-ascription"/>
```

A specific example of annotating the multimodal corpus is discussed in Section 3.3.

3.2.2 The GeM-Tools

The GeM-Tools are a set of computational tools developed by Hiippala (2015a). They afford the annotation of multimodal corpora drawing on the GeM model[3]; they furthermore allow for visualizing the annotations made. The GeM-Tools are written in Python, run as Jupyter notebooks[4] and use GraphViz[5] for visualizations. Python, as a computer programming language, is a script language which contains features of interpretability, compilation, interactivity, and object-orientedness. The Jupyter Notebooks used here are intended for "visualizing discourse structures in multimodal documents, as described using Rhetorical Structure Theory and annotated using scheme proposed in the Genre and Multimodality model" (Hiippala 2015a). GraphViz is a software used for the visualization of annotations. After annotating the three layers based on the GeM annotation scheme and storing the description into XML files, we use the GeM-Tools to display the XML data.

The GeM-Tools include three Jupyter notebooks which are intended to visualize (1) rhetorical structures, (2) layout structures, and (3) rhetorical and layout structures. If we want to visualize each of the above three structures, we need to run the specific categories of XML files. For instance, if we want to visualize the rhetorical structure, we open Jupyter notebooks, select the target item, and run the 'visualize_rst.ipynb file' in GeM-Tools. In order to obtain the graph of the rhetorical structure, we need valid XML files for both base and RST layers. The content was retrieved from the base layer before extracting the rhetorical relations between the content, as defined in the RST annotation. As a next step, the base and RST files need to be named before parsing/running them. The layout structure, and rhetorical and layout structure are visualized in a similar way. The XML file for the

[3] Visit http://www.purl.org/net/gem to find more about the GeM project.

[4] Jupyter notebooks is free and available at http://jupyter.org/.

[5] As GraphViz does not support Chinese by default, one needs to specify the font before the image output to avoid displaying unreadable codes. GraphViz is available at: http://www.graphviz.org/.

layout layer can visualize the layout structure. Similarly, if we run the XML files for the base, RST and layout layers together, we obtain a combined graph of the rhetorical and layout structure.

3.3 Annotating the Multimodal Corpus

3.3.1 Identifying Base Units

The first step of multimodal corpus annotation according to the original GeM scheme is to identify the base units since they constitute the smallest 'building blocks' or units on which other analytical layers draw. These minimal elements "serve as the common denominator for interpretative and textual elements as well as for layout elements" in the analysis of multimodal artefacts (Bateman 2008: 110). Basically, base units should not be larger than units described in other analytical layers.

Tab. 3: Recognized Base Units in Environment Protection Posters in CEPP.

Recognized Base Units in CEPP		
sentences	headlines	images
icons	logos/emblems	signs/symbols
list/sequence items	list/sequence labels	emphasized text
QR codes	hashtags	footnote label
publication/code numbers	publication/printing time	links to websites

- photographs, illustrations, cartoons, drawings, diagrams, charts, maps, etc. (without caption)
- captions of photographs, illustrations, cartoons, drawings, diagrams, charts, maps, etc.
- carrier-attribute: bubble, water drop, callout, text label, circles/pies, puzzle pieces, etc.
- text in photographs, illustrations, cartoons, drawings, charts, maps, etc., and carriers
- sentence fragments initiating a list
- footnotes (without footnote label)
- horizontal/vertical lines (solid, dotted, dashed) as delimiter between items
- lines which emphasize text (e.g., underline)
- lines which connect other units
- some punctuation marks such as highlighted question marks, exclamation marks, etc.
- copyright symbols, registered trademark symbols
- highlighted markers, arrows, etc.

The principles we need to follow are relative atomicity and objectivity. The principle of atomicity allows us to treat all text elements, even a single word as a base unit.

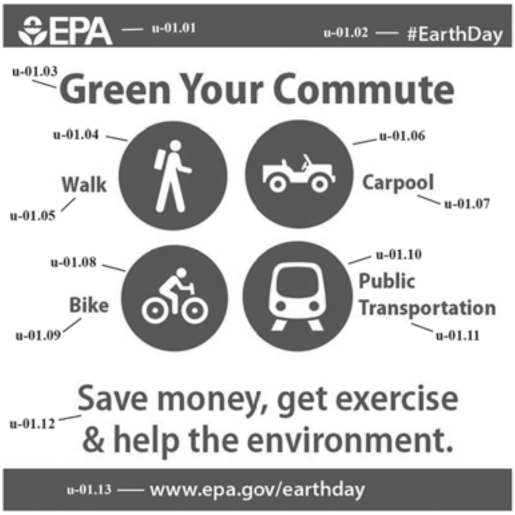
In order to avoid an explosion of base units, however, single words are usually not defined as minimal elements in most research (Zhang 2018). Generally, we intend to identify base units from the perspective of function, genre, and specific research objectives. In the GeM model as applied here, we consequently regard orthographic sentences as the minimal linguistic unit during analysis. As for the principle of objectivity, this requires the analyst to label all the basic elements precisely and comprehensively, without combining any basic units prematurely. For multimodal artifacts, the base units are open and non-exhaustive. Based on the base unit identification by Bateman (2008: 111) and Zhang (2018: 147), we list the adopted base units for environment protection posters used in my research in Table 3.

3.3.2 The GeM Base

After identifying the base units, we give an example here of a complete XML annotation of environment protection posters. We take the "Green Your Commute" poster as example because its simple design offers a good start for discussing the GeM annotation process. As shown in Figure 3, we first identify the base units of the poster US-1, label them and annotate the base layer. In the decomposing process, altogether 13 base units were obtained; they are given in Table 4. As it is labelled and annotated, six visual elements (logos[6] and illustrations) and seven verbal elements form the basic constitutes of the poster.

Table 4 gives all base units, distinguishing verbal units (non-bold) from visual units (bold): an EPA logo (u-01.01) is at the top left of the poster, and a hashtag (u-01.02) at the top right. A title (u-01.03), captions relating to four means of transportation (u-01.05, u-01.07, u-01.09, u-01.11) with their corresponding illustrations (u-01.04, u-01.06, u-01.08, u-01.10), and the results (u-01.12) of greening one's commute occupy the central part of the poster. At the bottom is a link to the institution's website (u-01.13), which enables readers to access more information.

[6] Although both verbal and visual elements are included in a logo, we still regard it as a single united visual unit for the purposes of the current analysis (but see, for example, Johannessen et al., this volume, for analyses of logos themselves).



Green Your Commute
US-1

Fig. 3: Example of labelling the base units.

Tab. 4: Labelling the base units of US-1.

label	unit
u-01.01	**Logo: EPA**
u-01.02	**Hashtag: #EarthDay**
u-01.03	Green Your Commute
u-01.04	**Illustration: A person walking**
u-01.05	Walk
u-01.06	**Illustration: A Carpool**
u-01.07	Carpool
u-01.08	**Illustration: A person riding a bike**
u-01.09	Bike
u-01.10	**Illustration: A train**
u-01.11	Public Transportation
u-01.12	Save money, get exercise, help the environment.
u-01.13	www.epa.gov/earthday

Based on the GeM annotation scheme and the GeM-Tools discussed above, the complete XML annotation of the base layer of US-1 is then written as follows, including the `alt` attributes to remind the analysts of its content:

```
<?xml version=”1.0” encoding=”UTF-8”?>
<gemBase>
<unit id=”u-01.01” alt=”Logo: EPA”/>
<unit id=”u-01.02” alt=”Hashtag: \#EarthDay”/>
<unit id=”u-01.03”> Green Your Commute</unit>
<unit id=”u-01.04” alt=”Illustration: A person walking”/>
<unit id=”u-01.05”> Walk</unit>
<unit id=”u-01.06” alt=”Illustration: A Carpool”/>
<unit id=”u-01.07”> Carpool</unit>
<unit id=”u-01.08” alt=”Illustration: A person riding a bike”/>
<unit id=”u-01.09”>Bike</unit>
<unit id=”u-01.10” alt=”Illustration: A train”/>
<unit id=”u-01.11”> Public Transportation</unit>
<unit id=”u-01.12”> Save money, get exercise, help the environment. </unit>
<unit id=”u-01.13”> www.epa.gov/earthday</unit>
</gemBase>
```

3.3.3 The Layout Structure

As explained above, the layout base consists of three parts: layout segmentation, realization information, and layout organization (the area model). Each layout unit specified in the layout segmentation necessarily has a visual realization. The GeM annotation scheme of the layout base is mainly suitable for grid-based documents. The layout structure of the environment protection posters is not the main focus of the present study, and we thus estimate the hierarchical structure of the poster in the GeM framework. The graphical representation of the corresponding layout structure of poster US-1 is shown in Figure 4. From the hierarchical organization of the graph, we can see that the poster can be divided into five parts: the header, the headline, the body, the result and the website, with the header containing a logo and a hashtag and the body containing four pairs of image and caption.

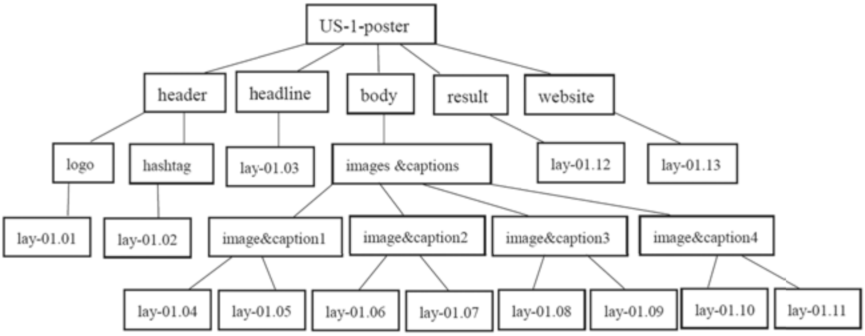


Fig. 4: A graph of the layout structure of the poster US-1.

3.3.4 The Rhetorical Structure

In order to illustrate the analyses of the rhetorical structure of environment protection posters, we draw again on the US-1 poster. As the logo of EPA is not marked as an RST segment, US-1 has altogether twelve RST segments including the hashtag (s-01.01), the title (s-01.02), the four parallel images and captions part (from s-01.03 to s-01.10), the result sentence part (s-01.11), and the link (s-01.12). All RST segments are marked as `<segment>`, so these segments can be recognized by the GeM-Tools. The tag `<span>` is used for asymmetric relations, whereas `<multi-span>` is used for symmetric relations with two nuclei. The multinuclear relation and intra-clausal relations are marked as `<multi-span>` and `<mini-span>` respectively. To make this clear, the XML annotation for the rhetorical layer of US-1 is also presented below. After the segmentation, all twelve RST segments form four spans and five multi-spans. Out of five multi-spans, four instantiate the relation of Multinuclear Restatement and the other one instantiates the relation of Joint. The four spans are nucleus-satellite relations. Based on the labeled “`id`”, “`xref`” and “`alt`”, it can be identified which segment forms a particular relation, and how they relate to the base level, that is, which base units form a respective segment. Afterwards, the specified verbal and visual elements of the base units can be identified. One multi-span or span may become the nuclei or nucleus and satellites of another multi-span or span, which leads to a hierarchical or recursive rhetorical organization.

```
<?xml version=”1.0” encoding=”UTF-8”?>
<gemRst>
<segmentation>
<segment id=”s-01.01” xref=”u-01.02” alt=”Hashtag: \#EarthDay”/>
<segment id=”s-01.02” xref=”u-01.03”> Green Your Commute</segment>
```

```
<segment id=”s-01.03” xref=”u-01.04” alt=”Illustration: A person
 walking”/>
<segment id=”s-01.04” xref=”u-01.05”> Walk</segment>
<segment id=”s-01.05” xref=”u-01.06” alt=”Illustration: A Carpool”/>
<segment id=”s-01.06” xref=”u-01.07”> Carpool</segment>
<segment id=”s-01.07” xref=”u-01.08” alt=”Illustration: A person
riding a bike”/>
<segment id=”s-01.08” xref=”u-01.09”> Bike</segment>
<segment id=”s-01.09” xref=”u-01.10” alt=”Illustration: A train”/>
<segment id=”s-01.10” xref=”u-01.11”> Public Transportation</segment>
<segment id=”s-01.11” xref=”u-01.12”> Save money, get exercise,
help the environment. </segment>
<segment id=”s-01.12” xref=”u-01.13”> www.epa.gov/earthday</segment>
</segmentation>
<rst-structure>
<span id=”span-01.01” nucleus=”s-01.01” satellites=”span-01.03”
relation=”enablement”/>
<span id=”span-01.02” nucleus=”s-01.12” satellites=”span-01.03”
relation=”enablement”/>
<span id=”span-01.03” nucleus=”s-01.11” satellites=”span-01.04”
relation=”result”/>
<span id=”span-01.04” nucleus=”s-01.02” satellites=”span-01.05”
relation=”solutionhood”/>
<multi-span id=”span-01.05” nuclei=”span-01.06 span-01.07 span-01.08
span-01.09” relation=”joint”/>
<multi-span id=”span-01.06” nuclei=”s-01.09 s-01.10”
relation=”multinuclear restatement”/>
<multi-span id=”span-01.07” nuclei=”s-01.07 s-01.08”
relation=”multinuclear restatement”/>
<multi-span id=”span-01.08” nuclei=”s-01.05 s-01.06”
relation=”multinuclear restatement”/>
<multi-span id=”span-01.09” nuclei=”s-01.03 s-01.04”
relation=”multinuclear restatement”/>
</rst-structure>
</gemRst>
```

Based on the valid XML files of both base and rhetorical layers of the poster US-1, the GeM-Tools (Hiippala 2015a) are then be able to visualize the rhetorical structure of the poster. The hierarchical structure of the rhetorical relations for the poster US-1 is shown in Figure 5. The body segment of the poster US-1 is the four means of transportation, which jointly function as solutions to green a commute. Therefore, the relation that holds between the title “Green Your Commute” and the joint four transportation means is Solutionhood. As for the four means of transportation, the relation between the four verbal texts and their corresponding visual images is Multinuclear restatement. All the solutions for greening one’s commute lead to the result of “Save money, get exercise and help the environment”. The hashtag and the link of the website are two ways to enable the public to obtain more relevant information and so are classified rhetorically as Enablement.

4 Analyzing the Multimodal Corpus

4.1 Analyzing Texts and Images

Table 5 presents us with a general picture of the number of base units, verbal units, and visual units realized in each poster in the corpus. As for the total number of base units, American posters (118) contain a slightly higher number of units, compared to the Chinese posters (99). The distribution of verbal and visual units in each sub-corpus is very similar: In CEPP-CN, the percentage of verbal and visual units is nearly even (verbal: 49.50%; visual: 50.50%). In CEPP-US, there are slightly more visual units than verbal ones (verbal: 47.46%; visual: 52.54%). It becomes apparent, as well, that poster CN-5 differs from the other posters in that the verbal and visual units are highly unequal in number. This poster will thus be discussed further below.

Tab. 5: Base units, verbal units, and visual units in each poster.

	Base units	Verbal	Visual		Base units	Verbal	Visual
CN1	14	6	8	US1	13	7	6
CN2	10	7	3	US2	9	3	6
CN3	33	23	10	US3	19	11	8
CN4	24	11	13	US4	50	20	30
CN5	18	2	16	US5	27	15	12
Total	99	49	50	Total	118	56	62
%	100%	49.50%	50.50%	%	100%	47.46%	52.54%

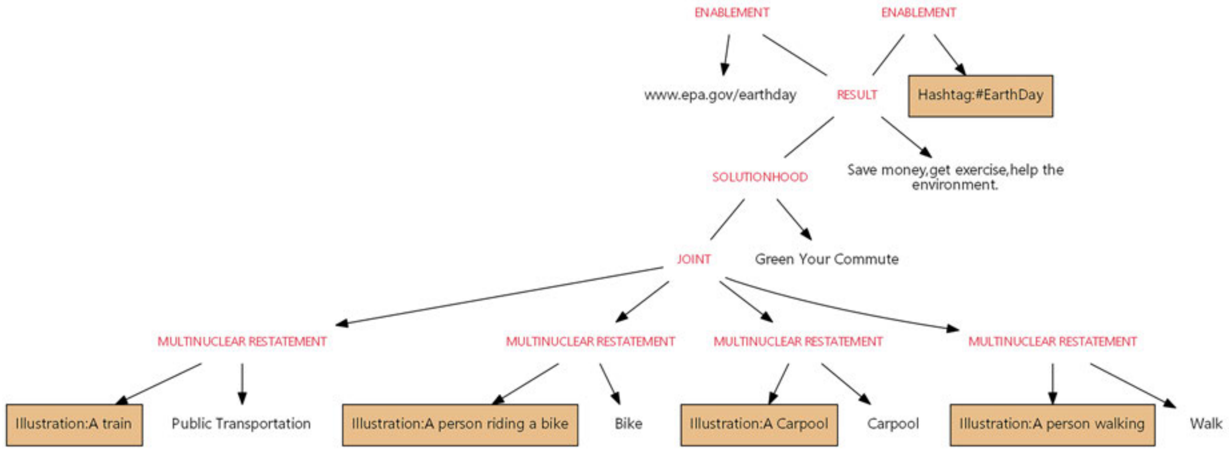


Fig. 5: Hierarchical structure of rhetorical relations for US-1.

4.2 Basic Statistics of Rhetorical Relations

In this section, we focus on the rhetorical relations used in the environment protection posters. All types and numbers of the three categories of rhetorical relations, namely, nucleus-satellite relations, multinuclear relations and intra-clausal relations, are shown respectively in Table 6, Table 7, and Table 8.

4.2.1 Nucleus-Satellite Relations

Table 6 presents twelve mononuclear relations used in the present corpus and their frequency of occurrence. Nearly half of the total number of the rhetorical relations, including Elaboration, Enablement, Motivation, Circumstance, and Result, occurs in both of the sub-corpora. However, relations of Manner, Background, and Purpose, and relations of Solutionhood, Interpretation, and Unconditional appear only in CEPP-CN and CEPP-US respectively. For the two sub-corpora, Elaboration is found to be the most frequent relation. This finding thus mirrors similar results from other studies. Elaboration is frequently used in many other genres, such as tourist brochures (Hiippala 2015a: 150) and public health posters (Zhang 2018: 177). In the present corpus, Elaboration takes up nearly half of the total amount among all the relations used in CEPP-CN. Through this comparative analysis, we find Chinese environment protection posters appear to attach importance to the explanation and background information of the poster since Elaboration, Circumstance, and Background relations are frequently used, whereas the American environment protection posters stress actually performing the action or taking measures as Elaboration, Enablement, and Solutionhood are frequently employed. By Fisher's Exact Test, there is a significant difference between the uses of rhetorical relations in the CEPP-CN and CEPP-US posters for the values in Table 6 (P=0.001159).

4.2.2 Multinuclear Relations

The data in Table 7 shows six multinuclear relations in the corpus. It is evident that the relation Multinuclear restatement occurs most frequently in CEPP-US with the number of CEPP-US almost seven times as many as in CEPP-CN. For posters in CEPP-US, the occurrence of the Multinuclear restatements relation is related to image-text relations, which concern images and texts restating each other specifically (e.g., "Image of a person walking" and "Walk" in US-1). As for posters in CEPP-CN, Multinuclear restatements occur only in relation to different languages. Both the relation between the Chinese texts and their corresponding English translations

Tab. 6: Nucleus-satellite relations in CEPP.

Nucleus-satellite relations	CEPP-CN	CEPP-US	Total
Elaboration	13	8	21
Enablement	1	5	6
Motivation	1	2	3
Circumstance	4	1	5
Result	2	2	4
Manner	4	0	4
Background	3	0	3
Purpose	3	0	3
Solutionhood	0	5	5
Interpretation	0	3	3
Unconditional	0	1	1
Total	31	27	58

Tab. 7: Multinuclear relations in CEPP.

Multinuclear relations	CEPP-CN	CEPP-US	Total
Multinuclear Restatement	3	26	29
Joint	7	5	13
Contrast	1	0	1
List	0	2	2
Addition	0	3	3
Total	11	36	47

(e.g., “绿色生活” and “Green Life” in CN-2) and the relation between Chinese characters and their corresponding Chinese *pinyin* (e.g., “第二十二届世界水日” and “DI-ERSHI’ER JIE SHIJIE SHUIRI” in CN-3) are Multinuclear restatements. The Joint relation has numerous occurrences in the CEPP-CN sub-corpus, amounting to more than half of the multinuclear relations. Chinese environment protection posters prefer to generate meaning by weaving images or images and texts together, which makes the Joint relation appear frequently. The rhetorical relation of Contrast appears only in CEPP-CN, and rhetorical relation of Addition only appear in CEPP-US. Again, by Fisher’s Exact Test, there is a significant difference between CEPP-CN and CEPP-US posters for the values in Table 7 (P=0.008905).

Tab. 8: Intra-clausal relations in CEPP.

Intra-clausal relation	CEPP-CN	CEPP-US	Total
Identification	0	10	10

4.2.3 Intra-Clausal Relations

Intra-clausal relations are provided in Table 8. Only one type of intra-clausal relation is found in the present corpus, which appears in US-4. The relation of Identification is used ten times in the recycling poster, which lists ten kinds of goods (US-4). The relation between the identifier label number and the identified image is Identification. The rhetorical relations holding between all the ten labeled illustrations and the correspondent verbal texts are all Multinuclear restatement relations.

4.3 Rhetorical Relations and Rhetorical Structures

Rhetorical relations include relations holding between textual segments and other semiotic resources. Table 9 presents the occurrences of Elaboration and Multinuclear restatement between texts and text-image. Elaboration occurs between verbal segments as well as verbal and visual segments. The theme as the nucleus and the content of theme "Live a green life" as the satellite instantiates the Elaboration relation holding between two verbal segments, as shown in Figure 6(a). Figure 6(b) shows the relation of Elaboration holding between verbal and visual segments, as "Choose public transportation" is the nucleus and different means of public transportation is the satellite. Multinuclear restatements in CEPP-CN (3) are produced solely between textual elements, which is reflected in two languages (e.g., CN-2) or two forms of one language (Chinese characters and *pinyin*) (e.g., CN-3). 26 occurrences of Multinuclear restatement in CEPP-US appear solely between text and image, as there are many cases in which captions are always presented with their corresponding visual displays to restate the information (e.g., US-1 & US-4 & US-5).

In environment protection posters, texts and images can be used separately as well as together to generate meaning and facilitate understanding. The rhetorical relations of List and Contrast in CN-5 and CN-2 respectively show how images work efficiently to make meaning. In CN-5, twelve paper cut animals are listed in a circle in a fixed order to form the Chinese zodiac, so the relation between the twelve paper

Tab. 9: Occurrences of elaboration and multinuclear restatement.

Rhetorical relations	Types	CEPP-CN	CEPP-US
Elaboration	texts	2	0
	text-image	11	8
Multinuclear restatement	texts	3	0
	text-image	0	26

cut images are classified rhetorically as List, as shown in Figure 6(c)[7]. In CN-2, the two visual segments, half a tungsten bulb and half a spiral-shaped energy-saving bulb, are comparable, which forms a Contrast relation, as shown in Figure 6(d).

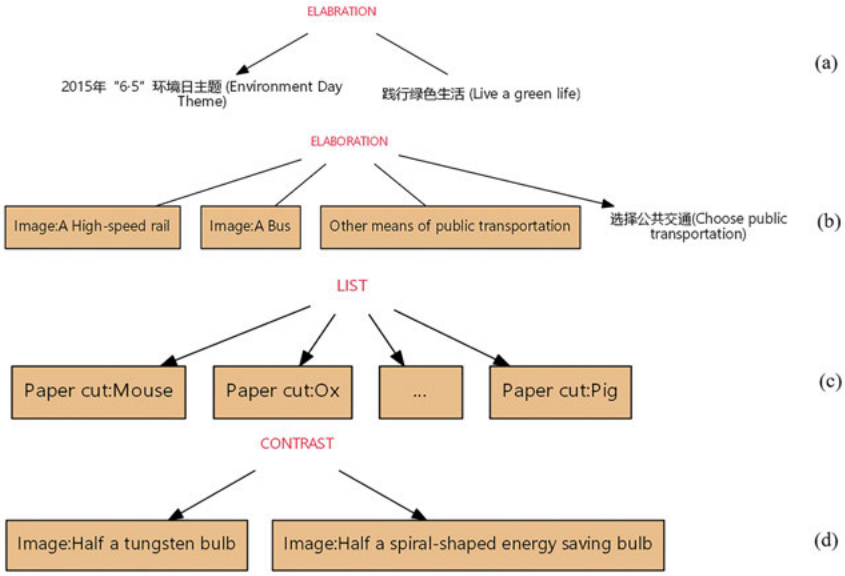

ELABRATION
2015年 "6·5" 环境日主题 (Environment Day Theme)
践行绿色生活 (Live a green life)
(a)
ELABORATION
Image:A High-speed rail
Image:A Bus
Other means of public transportation
选择公共交通(Choose public transportation)
(b)
LIST
Paper cut:Mouse
Paper cut:Ox
...
Paper cut:Pig
(c)
CONTRAST
Image:Half a tungsten bulb
Image:Half a spiral-shaped energy saving bulb
(d)

Fig. 6: Examples of rhetorical structures in CEPP.

[7] An abbreviated version is used in order to make the graph clearer.

5 Cross-Cultural Comparison

There are already many studies on how language and other meaning-making resources are culturally situated (Bowcher 2012) and how different semiotic resources are analyzed in broader contexts of culture (Bateman & Delin 2003; Thomas 2009; Kong 2013; O'Halloran 2014; Nekić 2015; Hiippala 2015b; Zhang 2018). Cross-cultural multimodal comparison studies include research on multimodal politeness (Idemaru et al. 2019), encoding and decoding of emotional speech (Li 2015), gestures (Kendon 2004; Caridakis et al. 2012; Urakami 2014), and speech and gestures (Aboudan & Beattie 1996; Gogate et al. 2015; Lin 2017), and so on. This study examines the Chinese and American environment protection posters, compares and contrasts the design of the documents from the two countries to support possible explorations of their cultural differences; two illustrative areas of such comparison are now presented.

5.1 Variation in Language and Typography

One obvious difference between Chinese and American environment protection posters analyzed is that the former are bilingual, while the latter are predominately monolingual. Table 10 shows that for Chinese posters, there are three categories: Chinese characters only, Chinese characters with *pinyin*, and Chinese and English bilingual version; while for the present corpus of American environment protection posters, only an English version is employed. Typography plays an important role in studies involving multimodal comparison across languages. Chinese and English belong to different writing systems, the former is logographic, while the latter is alphabetic. In the present study, vertical layout is much more common in Chinese characters (CN-3). For English, if the text is set vertically without changing the orientation of letters, it is often considered "inherently anti-typographic" (Kane 2002: 73).

5.2 Cross-Cultural Meaning-Making

In Chinese environment protection posters, images are joint together to convey meaning. For instance, in CN-5, the twelve paper cut animals are listed to represent the twelve Chinese zodiac signs and stand for all the species in the world. If we unite the twelve paper cut animals in a circle and the pair of scissors in the middle as a whole, a clock-shaped image is created, which symbolizes the reduced species with the passage of time. In addition, the homophonic pun strategy is employed

Tab. 10: Language version of the environment protection posters.

	Language version	Number of posters	Total posters
CEPP-CN	Chinese character	1	5
	Chinese character and *pinyin*	1	
	Chinese character and English	3	
CEPP-US	English	5	5

in CN-5. Specifically, the pronunciation of the first character of "剪刀" (scissors) in Chinese is "*jian*", which has the same sound as the pronunciation of the Chinese character "减" (reduced). Therefore, "剪少的生命" means the shortened lifespan or the reduced lives of the animals. American environment protection posters tend to demonstrate the Motivation in the posters directly. For example, in US-5, the rhetorical relation of Multinuclear restatement is used repeatedly between texts and images to present the endangered animals explicitly. Then it clearly calls for action from the public: "The world's threatened animals need you to remember!".

6 Discussion and Conclusion

The present study has explored how semiotic resources (languages and images) work together to generate meaning in environment protection posters. We summarize the main findings from the following three perspectives.

First of all, images are often used to restate the meaning of the verbal language. A visual image can re-express the meaning of a text, or a text line can be regarded as a caption to illustrate or interpret what the image presents. For instance, the texts in US-1 and the images near the texts restate the same messages. Similarly, the images and their associated names or titles placed near the images are re-expressions of each other.

Second, images are often used to elaborate the information presented through verbal language in order to get a better understanding of the poster. For example, in poster CN-1, the specific means of public transportation the images present (by train, by bus, and so on) are exemplifications of means of public transportation, which elaborates the title of the poster "Choices of Public Transportation". In the same way, in the recycling poster US-4, a visual viewing screen exemplifies the information made by the verbal text "Electronics".

Third, the American environment protection posters examined build on the rhetorical relations of Enablement and Solutionhood. Specifically, in the present corpus of CEPP-US, practical actions and measures are presented on how to protect the environment. For instance, according to three different situations, three correspondent methods are listed in poster US-3 to prevent the polluted runoff. Taking the recycling poster US-4 again as an example, the solutions are presented from three parts (Top 10 in the bin, 'Also recyclable' and the link of the website) in order to answer the question of the title "What can I recycle?".

To sum up, there are similarities and differences in the two sub-corpora CEPP-CN and CEPP-US at different levels. However, the corpora in the present study consist of just ten environment protection posters, a limited number including only environmental posters, not involving posters in other fields. A large number of randomly sampled posters need to be analyzed and compared by means of statistical analyses. Therefore, in future work we will try to expand the corpus and do further comparative studies to analyze the language, image, layout, and rhetorical relations participating in meaning making to generate effective environment protection posters and educate the public more generally. The possibility and applicability of this research paradigm for the analysis of other subjects and genre materials will be explored in the future as well, but we believe the discussion in the present chapter has already indicated beneficial directions and methods for these research goals.

Acknowledgements

This work was supported by the PhD Candidate Overseas Visiting Program, the PhD Candidate Key Research Program of Xi'an International Studies University under research grant syjsb201701, and China National Foundation for Social Sciences under research grant 20BYY221. Thanks to John Bateman for his guidance, revisions, and constructive comments of this chapter.

Bibliography

Aboudan, Rima & G. Beattie. 1996. Cross-Cultural Similarities in Gestures: The Deep Relationship between Gestures and Speech which Transcends Language Barriers. *Semiotica* 111. 269–294. https://doi.org/10.1515/semi.1996.111.3-4.269.

Bateman, John A. 2008. *Multimodality and Genre: A Foundation for the Systematic Analysis of Multimodal Documents*. Basingstoke: Palgrave Macmillan.

Bateman, John A. & J. L. Delin. 2003. Genre and Multimodality: Expanding the Context for Comparison across Languages. In D. Willems, B. Defrancq, T. Colleman & D. Noël (eds.), *Contrastive analysis in language: identifying linguistic units of comparison*, 230–266. Houndsmill: Palgrave Macmillan.

Bateman, John A. & J. L. Delin. 2006. Rhetorical Structure Theory. In K. Brown (ed.), *The Encyclopedia of Language and Linguistics*, vol. 10, 588–596. Amsterdam: Elsevier 2nd edn.

Bateman, John A., J. Wildfeuer & T. Hiippala. 2017. *Multimodality – Foundations, Research and Analysis. A Problem-Oriented Introduction*. Berlin: De Gruyter Mouton.

Bowcher, Wendy. 2012. *Multimodal Texts from Around the World: Cultural and Linguistic Insights*. Basingstoke: Palgrave Macmillan.

Caridakis, George, J. Wagner, A. Raouzaiou, F. Lingenfelser, K. Karpouzis & E. Andre. 2012. A Cross-Cultural, Multimodal, Affective Corpus for Gesture Expressivity Analysis. *Journal on Multimodal User Interfaces* 7. 121–134. https://doi.org/10.1007/s12193-012-0112-x.

Gogate, Lakshmi, M. Maganti & L. Bahrick. 2015. Cross-Cultural Evidence for Multimodal Motherese: Asian-Indian Mothers' Adaptive Use of Synchronous Words and Gestures. *Journal of Experimental Child Psychology* 129. 110–126. https://doi.org/10.1016/j.jecp.2014.09.002.

Halliday, Michael A. K. & C. M. I. M. Matthiessen. 2013. *Halliday's Introduction to Functional Grammar*. London and New York: Routledge 4th edn.

Hiippala, Tuomo. 2014. Multimodal Genre Analysis. In S. Norris & C. D. Maier (eds.), *Interactions, Images and Texts: A Reader in Multimodality*, 111–123. Berlin: De Gruyter Mouton.

Hiippala, Tuomo. 2015a. Gem-Tools: Tools for Working with Multimodal Corpora Annotated Using the Genre and Multimodality Model https://doi.org/10.5281/zenodo.33775.

Hiippala, Tuomo. 2015b. *The Structure of Multimodal Documents: An Empirical Approach*. London: Routledge.

Hiippala, Tuomo. 2017. An Overview of Research within the *Genre and Multimodality* Framework. *Discourse, Context and Media* 20. 276–284. https://doi.org/10.1016/j.dcm.2017.05.004.

Idemaru, Kaori, B. Winter & L. Brown. 2019. Cross-Cultural Multimodal Politeness: The Phonetics of Japanese Deferential Speech in Comparison to Korean. *Intercultural Pragmatics* 16(5). 517–555.

Kane, John. 2002. *A Type Primer*. London: Lawrence Kind Publishing.

Kendon, Adam. 2004. *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Kong, Kenneth C.C. 2013. A Corpus-Based Study in Comparing the Multimodality of Chinese- and English-Language Newspapers. *Visual Communication* 12(2). 173–196.

Li, Aijun. 2015. *Encoding and Decoding of Emotional Speech: A Cross-Cultural and Multimodal Study between Chinese and Japanese*. Springer. https://doi.org/10.1007/978-3-662-47691-8.

Lin, Yen-Liang. 2017. Co-Occurrence of Speech and Gestures: A Multimodal Corpus Linguistic Approach to Intercultural Interaction. *Journal of Pragmatics* 117. 155–167. https://doi.org/10.1016/j.pragma.2017.06.014.

Mann, William & C. Matthiessen. 1991. Functions of Language in Two Frameworks. *Word* 42(3). 231–249. https://doi.org/10.1080/00437956.1991.11435839.

Mann, William C. & S. A. Thompson. 1987. Rhetorical Structure Theory: A Theory of Text Organization. Tech. Rep. RS-87-190 USC/Information Sciences Institute. Reprint series.

Mann, William C. & S. A. Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text* 8(3). 243–281.

Matthiessen, Christian & K. Teruya. 2015. Grammatical Realizations of Rhetorical Relations in Different Registers. *Word* 61(3). 232–281. https://doi.org/10.1080/00437956.2015.1071963.

Matthiessen, C.M.I.M. 2002. *Combining Clauses into Clause Complexes: A Multi-faceted View* 237–322. John Benjamins. https://doi.org/10.1075/z.110.13mat.

Matthiessen, C.M.I.M. 2014. In Developing Systemic Functional Linguistics: Theory and Application. *Appliable Discourse Analysis* 138–208. London: Equinox.

Nekić, M. 2015. *Tourist Activities in Multimodal Texts: An Analysis of Croatian and Scottish Tourism Websites*. London: Palgrave Macmillan. https://doi.org/10.1057/9781137397911.

O'Halloran, Kay L. 2014. Multimodal Discourse Analysis. In K. Hyland & B. Paltridge (eds.), *The Bloomsbury Companion to Discourse Analayis*, 120–137. London and New York: Bloomsbury.

Taboada, María Teresa & W. C. Mann. 2006. Applications of Rhetorical Structure Theory. *Discourse Studies* 8(4). 567–588.

Thomas, Martin. 2009. *Localizing Pack Messages: A Framework for Corpus-Based Cross-Cultural Multimodal Analysis*. Leeds: Centre for Translation Studies, University of Leeds PhD Thesis. http://corpus.leeds.ac.uk/~martin/thesis/martin_thomas_thesis_2009_semi-skimmed.pdf (last accessed: 1 September 2021).

Urakami, Jacqueline. 2014. Cross-Cultural Comparison of Hand Gestures of Japanese and Germans for Tabletop Systems. *Computers in Human Behavior* 40. 180—189. https://doi.org/10.1016/j.chb.2014.08.010.

Zhang, Peijia. 2018. *Public Health Education Through Posters in Two World Cities: A Multimodal Corpus-Based Analysis*. University of Hong Kong dissertation.

Loli Kim and Jieun Kiaer

Conventions in How Korean Films Mean

A Pilot Testing ‘Segmented Film Discourse Representation Structures’

Abstract: The concept of ‘conventions’ in how meaning is made has become an implicit part of a dogma that surrounds South Korean films (K-film) popularized through the Korean Wave. However, the nature and content of any conventions discussed is only addressed vaguely and mostly in terms of narrative rather than film form. It also lacks much of the theoretical background and appropriate terminology needed to situate conventionalized usages within other relevant fields of film and communication research (e.g., multimodal film analysis, semiotics, semantics, pragmatics). In this chapter, we present a pilot study whose aim is to evaluate whether ‘Segmented Film Discourse Representation Structures’ (SFDRS) — a formally specified notion of film discourse — can help to identify conventions in meaning-making processes in K-films in a reliable fashion. This framework has not previously been employed for analyzing conventions, although it has been suggested. We focus our analysis specifically on the ‘final goodbye between parent/guardian and child’ event, extracted from films selected for their similar narrative and production criteria. By transferring film extracts to SFDRS and comparing the resulting representations across films, it was possible to identify certain reoccurring patterns in their configurations. We suggest that these stand as good candidates for conventions among these films. Further, patterns were found strictly within the boundary lines of meaning types at every level of discourse, supporting the validity of the patterns found being conventions in types of meaning and suggesting that these conventions may well be found in similar events in further films. Thus, the results of this study demonstrate that conventions benefit from more empirically-supported testing, and that SFDRS is a means of doing so.

Keywords: film language convention, Korean film, empirical multimodal film analysis, Segmented Film Discourse Representation Structures, Korean socio-pragmatic communication

1 Motivation

Conventions in film are essentially reoccurring ways of expressing particular kinds of meaning that establish themselves in a community of use. Such conventions

https://doi.org/10.1515/9783110725001-010

may involve anything from sharing single elements to groups of films coming to exhibit similar structures overall. For instance, if we consider events in film narrative to constitute types of meaning, then such meanings could be a ‘final goodbye’ between characters, the act of ‘revenge’, ‘shopping’, a person ‘starting the day’, or two people ‘falling in love’. Each of these events has a clearly distinguishable purpose and, in general, relatively well demarcated boundaries. They also commonly exhibit internal structure, constructed from a series of shorter happenings. Although narrative can be constructed in a multitude of ways, one can assume that the components that are employed and their logical organization will generally support comprehension. For instance, if a scene in a film is intended to show a person ‘starting the day’, then activities associated with this event (such as a person getting out of bed, taking a shower, and then eating breakfast) would need to be presented in such a way that recipients can construct an appropriate order of occurrence of events. One of the means of bringing about support for interpretation is precisely by the adoption of conventionalized associations between forms and intended kinds of meanings that audiences familiar with the conventions can recognize.

Similarly, ‘getting out of bed’, ‘taking a shower’, and ‘eating breakfast’ will also be composed at even finer levels of discourse by elements associated with those activities rather than others. For instance, ‘eating breakfast’ might include camera shots of the food, a person eating, drinking coffee, and reading a newspaper at the breakfast table, whereas ‘taking a shower’ might include shots of steam, water pouring from the shower head, and a person with bubbles in their hair. When aspects of these configurations are repeated in multiple films with the same types of meaning they are good evidence that conventionalized forms of expression are being employed. Filmic presentations of events then might become increasingly schematic because film-makers can rely on audiences filling the gaps by drawing on the meanings conventionally signaled. Particular shots and sequences of shots can then come to stand in for complex embedding narrative situations.

The term ‘convention’ can be found in many discussions that concern how meaning is made in film, including literature dedicated to contemporary South Korean film (K-film) popularized through the Korean Wave. However, the nature and content of any conventions discussed is only addressed vaguely and mostly in terms of narrative rather than film form. This then lacks much of the theoretical background and appropriate terminology that would be needed to situate conventionalized usages within other relevant fields of film and communication research (e.g., multimodal film analysis, semiotics, semantics, pragmatics). This situation notwithstanding, conventions have already become an implicit part of a K-film dogma despite the absence of clear definitions, focused and thorough investigation, and empirical analysis. In this chapter, we present a pilot study whose aim is to explore whether a formally specified notion of film discourse can help identify conventions in meaning-making processes in K-films in a reliable fashion.

To do this, we draw specifically on ‘Segmented Film Discourse Representation Structures’ (SFDRS) (Wildfeuer 2014) as a semantic tool for the formal description of how meaning is constructed multimodally in film. This framework has not previously been employed for analyzing conventions, although this possibility has been suggested (Wildfeuer 2014: 182). We develop this further here. SFDRS is the result of systematically describing how meaning is made logically in film. This approach to description makes the specific use of filmic resources and their interactions at finer and higher grained levels of discourse accessible within single overarching ‘filmic discourse structures’. These structures can be compared across film instances so that reoccurring patterns may be taken as indications of conventions at work. On this basis, therefore, SFDRS has the potential to be used to build empirically upon the limited descriptions provided for conventions in K-film so far.

Although there is a vast range of K-films to explore, we select for the purposes of our pilot study a small corpus of films with some readily recognizable reoccurring events; this corpus will be presented in Section 3.1. In particular, each film contains the narrative event of a ‘final goodbye between a parent/guardian and child’ and we extract the filmic construction of this event for focused analysis. In Section 3.2 we explain the process of transferring a filmically expressed narrative event to an SFDRS description using examples extracted from our analysis. Here we build on the examples of transference given in Wildfeuer’s original model of SFDRS (2014), extending the process further to divide higher and finer levels of discourse granularity more clearly by decomposing main activities (hereafter ‘higher activities’) into lower-level units of meaning. This achieves a clear separation of three levels of meaning-making processes — higher activities that compose the narrative event, segments of meaning that compose the higher activities, and the composition by which the meaning of each segment is inferred. Each of these levels of discourse provides a separate opportunity to compare samples and identify corresponding conventions. In Section 3.3 we consequently explain how we compare the instances of discourse structures captured in this way. In Section 4 we provide a specific description of the SFDRS of the narrative events described in Section 3.1 and, in Section 5 discuss our results. Finally, in Section 6 we discuss factors that affected the identification of conventions and propose future developments for SFDRS so as to capture meanings in K-film that are particularly embedded in Korean culture.

We argue that SFDRS and the development we pursue here will prove particularly important for future methodologies developed specifically for understanding K-film. In South Korea (Korea), verbal/non-verbal modes are used to indicate social hierarchy (according to age and rank, and potentially class and gender), interpersonal relations (e.g., intimacy and respect), mood and emotions, style, perspective, and attitude. In Korea socio-pragmatic meanings effect the forms and functions in every single sentence (Kiaer 2019) and, likewise, are mirrored in non-verbal utterances such as hand gestures, eye contact, posture, and traditions like bowing. Engaging with meaning-making of this kind has been suggested to demand a "much more fine-grained, formal framework to adequately capture the multi-dimensional, subjective, and socio-pragmatically sensitive meanings" (Kiaer 2020). Although we will not discuss these Korean socio-pragmatic meanings in this study, SFDRS was selected keeping in mind precisely the critical future developments necessary for understanding the complex interactional dynamics of K-film and beyond.

2 Theoretical Background

While meaning is a popular subject in film studies, it is rarely addressed formally to the extent that particular patterns of inference in interpretation can be followed. This is even more the case in K-film studies, where the discussion of narrative styles and the techniques employed by certain genres or regions of film remains overwhelmingly informal. There are, nevertheless, many discussions of conventions in K-film. Yecies & Shim (2015), for example, discuss narrative conventions in Korean melodramas. Ok (2009: 38) discusses a shift in the discourse of Korean national cinema to accommodate "Hollywood conventions" in narrative structure and spectacle and also "generic conventions" that were employed in genre films. Choe (2009) discusses the difference in conventions between *Old Boy* (2003) and the Bollywood remake *Zinda* (2006), explaining how conventions in violence and spectacle differ rather than film language as a whole. Choi (2010: 34, 40) describes conventions in K-film during the pre-Korean Wave period as sometimes "loose" and "inconsistent" and explains that the adoption of Hollywood conventions resulted in "tighter narratives, eliminating the episodic narratives previously characterizing Korean cinema". And, finally, Kim (2006) links the hybridization and Hollywoodization of conventions to viewer comprehension stating that:

> familiar stories based on collective memory, current cultural values and sentiments are enough for national audiences to interpret the Korean blockbuster as their own national cinema, to feel more attached towards it, and to share a certain level of emotions with other nationals in relation to it notwithstanding the fact that the typical Korean blockbuster contains a number of externally imported elements such as styles, spectacle dominating format, and so on. (Kim 2006: 6)

There have also been numerous appeals to notions of ‘logic’ in discussions of conventions in K-film and since some of the earliest writings. Primarily, this occurs through use of the term ‘coherence’, a term used in the same sense as it is defined by the Oxford English Dictionary (2020) to refer to conventions being ‘logical and consistent” . In one of the first essays written on K-film in English, Willemen (2002) recognizes the problematic meaning-making processes in films before the boom of international popularity at the end of the 1990s that marks the beginning of the Korean Wave (Dal 2016). In particular, he refers to the films of the 1970s and 1980s, and in doing so unintentionally describes periodical conventions in how meaning is made. Willemen describes these conventions as a problem in which “spatial relations in a scene operate differently, with the narrator and the viewer being inscribed into a scene in ways that the dominant Euro-American conventions of realist cinema were designed to displace”. He further states that “assumptions underpinning Western cinema’s notion of spatial and psychological coherence do not apply” to K-films (Willemen 2002: 7).

Another area of K-film study that can be linked to reoccurring patterns of meaning and processes of comprehension has explored the connections between changes in conventions and the increase of international popularity. They do this by referring to the employment of new ways of communicating that could be made sense of by wider audiences. For example, Han (2011: 64) discusses the language of Korean national cinema and the development of the Korean blockbuster via the language of Hollywood cinema. Also, Smith (2013: 188–189) refers to “universal or familiar styles” and “generic elements” being employed in certain genres and their potential role in the increase of K-films popularity at international film festivals.

The longevity of these ‘logical’ conventions in K-film is clearly indicative of the need both for a logical approach for their analysis and an appropriate empirical underpinning for reliable results. Former studies have not generally applied any logical terminology, however, and have similarly made little appeal to empirical methods. In the approach taken in this chapter, we seek in contrast to draw on explicit accounts of the logical organization of filmic discourse when analyzing films and narrative elements from the perspective of their role in forming conventions.

Logical approaches to understanding how meaning is made and understood in film on film’s terms have been advocated increasingly in recent years. Bateman & Schmidt (2012), for example, propose one thorough foundation for empirical film analysis. They incorporate multimodal theories with accounts that include discourse semantics in order to build methodologies that seek to identify the concrete details of film responsible for constructing discourse, but do not explicitly address convention. One approach that has explored conventions from a logical perspective is that of Cumming et al. (2017). This study takes a semantic perspective on a pair of conventions governing spatial relations between viewpoints. They argue that the semantic view provides the correct account of the means by which films express their content, and that in order to explain regularities of interpretation among viewers the semantic perspective is necessary. In conclusion they state that semantic explanations of conventions are “at least viable, by defusing the threat of counterexample, and even plausible, by contextualizing them within the broader framework of coherence relations” (Cumming et al. 2017: 27). And finally, as remarked above, Wildfeuer (2014) provides a framework developed from a dynamic semantic theory of discourse precisely for the systematic examination of film interpretation from a logical perspective. This framework was developed to provide a detailed description of how recipient’s meaning-making processes are guided by coherence and structure in film’s text.

We consider a logical orientation for exploring conventions to be particularly promising. This provides a ‘textual’ perspective on film analysis in which the search for explanations of how meaning in film is constructed and comprehended takes centre stage (Morey Hawkins 2018). Considerations of film ‘as text’ have a long history (cf. Bentley 1995) and the explicit adoption of a linguistically-motivated notion of discourse takes this considerably further. Crucial here is a perspective in which film is recognized in its ability to ‘fulfil communicative purposes that are in many ways analogous to those achieved by sequences of linguistic elements” (Bateman & Wildfeuer 2014: 180). If we consider film in such linguistic terms, then what is really being talked about when the term ‘convention’ is used (in its many facets) is the more systematic side of ‘film language’. In this respect, as Kim (2006: 7) argues, conventions “feed into the ways in which an audience responds to and interprets a film”. The extensive planning conducted in film production — from production design, art direction, and editing to scripted and directed dialogue and behavior — is under this view employed to lead viewers dynamically to realizations at certain moments. Viewers’ comprehension of this purposeful manipulation is strongly supported by conventionalized communicative strategies, which we aim here to uncover.

3 Method

As indicated above, our primary method for uncovering conventions will be to analyze selected film extracts in terms of their logical composition using Wildfeuer’s (2014) SFDRS framework. SFDRS was developed for film from a dynamic semantic framework called ‘Segmented Discourse Representation Theory’ (SDRT) (Asher & Lascarides 2003), in which the interaction between discourse coherence and discourse interpretation is drawn in a logically precise manner called ‘Segmented

Discourse Representation Structures' (SDRS). SDRT is a theory from the field of linguistics that builds upon DRT (Kamp & Reyle 1993; Kamp 1981) – "the most established model developed in the context of dynamic semantics" (Wildfeuer 2014: 39). It does so by enhancing the logical forms of discourse with rhetorical relations whose truth conditional effects contribute meaning beyond what is literally being said. It is this characteristic that also makes SDRT adaptable for the analysis of meaning in film because meaning is often implied rather than being made explicit. Likewise, it is this characteristic that makes SFDRS a promising approach for analyzing meaning specifically in K-film, since non-Korean researchers, particularly those of Anglophone and other Western European origins, run a considerable risk of misunderstanding the meaning-making processes involved due to the unfamiliar systems of socio-pragmatic meaning-making that are active (Kiaer et al. 2020; Kiaer 2020, 2019). Our goal in this section is to explain how the films of our sample were selected and subsequently to show how those segments were analyzed and transferred to SFDRS descriptions. We then set out the means by which the resulting SFDRS descriptions were compared in order to find patterns of reoccurrence that might stand as conventions.

3.1 Sample

The films selected as sources for the extracts that we analyze are listed in Table 1. In the table we show in addition the criteria that were employed for the selection. The aim of these criteria was to select film extracts with similar narrative and production criteria. Narrative criteria included: type of narrative event, the temporal position of the event in the film, characters involved, and genre. Production criteria included: release date, period of film production, and films were all Korean. Additionally, two of the films were chosen because they have the same director. The purpose of this was to explore conventions on a directorial microscale.

Despite the overlap in events and narrative types exhibited by the three films, they are each quite distinct in their storylines as well. The first and second films are taken from Park Chan-Wook's revenge trilogy. *Old Boy* (Romanisation: Oldeuboi, hangeul: 올드보이), the second in the trilogy, tells the story of Dae-su, a degenerate failure who is unlawfully imprisoned in a cell for 15 years. After his release he meets and falls in love with a young woman named Mi-do, who turns out to be his daughter. Dae-su undergoes hypnosis in order to forget that Mi-do is his daughter and prevent either of them from ever having to deal with the consequences of knowing the truth. The scene extracted for analysis is then the final scene, in which a metaphorical final goodbye between Dae-su and his daughter is realised. *Sympathy for Lady Vengeance* (Romanisation: Chinjeolhan Geumjassi

Tab. 1: An overview of samples and the narrative and production criteria for which they were selected.

Criteria	Old Boy (2003)	Sympathy for Lady Vengeance (2005)	The Man from Nowhere (2010)
Narrative event type	Goodbye between parent and child	Goodbye between parent and child	Goodbye between guardian and child
Event position	Final scene	Final scene	Final scene
Protagonists in event	Parent and child	Parent and child	Guardian and child
Time of extract	01:55:15–01:57:31 (when shot fades out before title credits)	01:47:23–01:51:32 (when shot cuts to title credits)	01:52:06–01:54:48 (when shot fades out before title credits)
Director	Park Chan-wook	Park Chan-wook	Lee Jeong-beom
Genre	Revenge genre	Revenge genre	Revenge genre
Release year/ film period	2003/Korean Wave period	2005/Korean Wave period	2010/Korean Wave period

'Kind-hearted Geumja', hangeul: 친절한 금자씨) is the third film in the trilogy and tells the story of the woman Geum-ja who is also released from prison after a mistaken imprisonment. The extracted scene, again the final scene of the film, sees Geum-ja and her long-estranged daughter, Jenny, reunited to say goodbye before Jenny returns to her adoptive parents. The third and final film selected is *The Man from Nowhere* (Romanisation: Ajeossi 'Mister', hangeul: 아저씨) directed by Lee Jeong-beom. This film tells the story of Tae-sik, a mysterious man who embarks on a bloody rampage when a neglected child who has befriended him, So-mi, is kidnapped by an organ trafficking drug ring. In the scene selected for analysis, Tae-sik and So-mi say their final goodbye before Tae-sik is arrested.

3.2 From Film to SFDRS

The process of transferring film to its SFDRS description can be divided into two stages. We will use examples from the process of transferring the 'goodbye between parent and child' event in *Old Boy* (2003) to SFDRS to demonstrate this.

In the first stage, film extracts are segmented. Current psychological research has shown segmentation to play a "critical role" in understanding in which "observers segment ongoing activity into meaningful events" (cf. Zacks et al. 2010). We employ segmentation for analytic purposes and, in general, one would aim for segmentations that would align with empirical studies of event segmentation as well. For current purposes, however, we approach this by first seeking narrative events with clear start and end points. In the current case, this role is served by the 'goodbye between parent/guardian and child' event. Demarcated events are then transcribed according to their temporal and spatial parts, again following Zacks et al. (2010: 1). For instance, a person arrives at a location, then meets another person, and then says something to that person. Each shift from one activity to another creates another temporal part and together they form the entire event. Zacks et al. (2010) define spatial parts as including the people and objects that form what is being observed, for example, a marketplace could be broken down into stalls, foods, and people. Commonly, when analyzing the finer grain levels of discourse, which largely concern spatial parts, shifts in spatiality are therefore more prevalent, while, at the higher level of discourse, shifts tend to be temporal going from one activity to another. The segmentation is then refined by transferring the rough transcription to 'logical forms'. These forms follow the general conventions for dynamic semantics of the formal tradition extended by Wildfeuer to include the formal, logical description of meaningful multimodal interactions. An example of a logical form from a 'meeting' event is shown in Table 2. The segments constituting the narrative event are represented by logical forms grouped within the standard box notation format in order to capture discourse segments and hierarchical relationships between discourse segments.

Tab. 2: The logical form employed in the analysis.

$\boldsymbol{H_{\pi_1}} = \boldsymbol{meet}$
$e_{\pi_{1a}} = arrive$
[v] Mi-do (a) [v] Enters shot (a.1) [a] Mi-do: "What's going on" (a.2) [v] Dae-su (b)
$a, a.1, a.2, b \mathrel{\vert\!\sim}$ **arrive** $(e_{\pi_{1a}})$

Following Wildfeuer (2014), the lower part of the box shown in the figure lists the various referents in the event, labelled alphabetically and also marked with [v] for visual or [a] for audio. In addition, we include numerical labels in the presentation of the logical form alongside the alphabetical ones to indicate the characters to which particular referents belong. This development was employed as an organizational measure because of the numerous types of visual and audio that can often be found within single segments. The bottom line of the box includes a statement of defeasibility; the labels show which referents are required in inferring the meaning. For example, in Table 2, the eventuality defeasibly inferred (as indicated by the defeasible eventuality symbol $\mid\!\sim$ over the referents labelled 'a, a.1, a.2, b' is 'arrive', which is itself labelled '$e_{\pi_{1a}}$' to indicate the segment's position in the discourse. $e_{\pi_{1a}} = arrive$ is then used as a title in the middle section of the logical form.

Since discourse within the formal discourse representation framework is considered to exhibit a hierarchical organization, this information must also be anchored into the discourse structure. Consequently, we articulated the presentation of the logical form further to specify the unit of discourse to which some logical form belongs within the larger narrative event under analysis. This allows us to explicitly represent how lower grain levels of discourse construct higher grain levels of discourse in order to achieve a broader picture of any conventions in meaning-making processes being employed. This therefore increases our opportunities to identify conventions. The introduction of a higher activity in an SFDRS is reflected in the outer box within which the logical form sits in Table 2. In the outer box, the higher activity label H_{π_1} is used to show that $e_{\pi_{1a}} = arrive$ is to be considered a part of the construction of the larger unit of discourse $H_{\pi_1} = meet$. We will see further examples and show how this is used below and particularly in section 4.

Stage two of the transcription process is concerned with the construction of hierarchical discourse structures by which each subsequent discourse segment is added dynamically into the growing discourse structure for the entire unit being analyzed. Discourse segments are added by examining which discourse relations apply. Discourse relations impose several logical constraints, both defeasible and non-defeasible, and these constraints must hold over the logical forms present in the discourse in order for the corresponding relations to be considered applicable. Thus, by inferring the various relations that hold between logical forms, a discourse's structure is created. The discourse relations relevant for film are defined by Wildfeuer (2014: 59) as follows: 1) *Narration* relation, 2) *Contrast* relation, 3) *Parallel* relation, 4) *Background* relation, 5) *Result* relation, 6) *Explanation* relation, and 7) *Elaboration* relation. As an example, for the discourse relation of *Narration* to apply, the logical segments involved must directly follow one another in sequence and their events must not overlap temporally. If the second discourse segment instead specifies further information about the first so that there is temporal overlap, then the Narrative relation cannot apply. In this case an *Elaboration* discourse relation would better account for the temporal consequence of information in one segment explaining more about something in the segment before it.

We can see this with respect to the current example segments in the analysis shown in Table 3 where, on the left-hand-side in H_{π_1}, $e_{\pi_{1a}} = arrive$ overlaps temporally with $e_{\pi_{1b}} = join$. The camera perspective switches from a birds-eye view upon a figure entering the camera frame to a close-up that reveals that Mi-do is joining Dae-su. We thus conclude that *Elaboration* is the most appropriate relation holding between the arrival and the joining events as shown in the figure. On the other hand, on the righthand side of the figure in H_{π_2}, there are strict temporal sequences between 'comfort', 'see', 'say', and 'look' ($e_{\pi_{2a}}$-$e_{\pi_{2d}}$) and so the *Narration* relation wins in each case.

It is also possible for two or more discourse relations to hold of the same discourse segments. This is because "coherence is not a yes/no matter but can vary in quality" (Asher & Lascarides 2007: 12) as is explained in Asher & Lascarides's (2003) principle of *Maximising Discourse Coherence* (MDC). In such cases, the strongest discourse relation must be determined by checking all the conditions given in the definitions of those relations. Further details are given in Wildfeuer (2014).

Tab. 3: A formal description of the SFDRS of the higher activities *Meet* and *Conversation* in the *Goodbye between parent and child* event from *Old Boy*.

π_0 = goodbye between parent and child

π_0:

$\pi_1 = H_{\pi_1}$

π_1:

$\pi_{1a}, \pi_{1b} = e_{\pi_{1a}}, e_{\pi_{1b}}$

π_{1a}, π_{1b}

$Elaboration(\pi_{1a}, \pi_{1b})$

$\pi_2 = H_{\pi_2}$

π_2:

$\pi_{2a}, \pi_{2b}, \pi_{2c}, \pi_{2d}, \pi_{2e} = e_{\pi_{2a}}, e_{\pi_{2b}}, e_{\pi_{2c}}, e_{\pi_{2d}}, e_{\pi_{2e}}$

π_{2a}, π_{2b}	$Narration(\pi_{2a}, \pi_{2b})$
π_{2b}, π_{2c}	$Narration(\pi_{2b}, \pi_{2c})$
π_{2c}, π_{2d}	$Narration(\pi_{2c}, \pi_{2d})$
π_{2d}, π_{2e}	$Elaboration(\pi_{2d}, \pi_{2e})$

π_1, π_2 $Narration(\pi_1, \pi_2)$

Once discourse relations have been applied between the logical forms, the SFDRS can be drawn to describe the complete discourse structure of the film extract. Together, logical forms and the SFDRS diagram provide descriptions of finer and higher grained levels of discourse within a single, uniform representation. Whereas

Wildfeuer's SFDRS representation was designed to capture logical relations describing the structure of segment eventualities composing a film extract, our representation articulates the presentation of the SFDRS further by also specifying the structure of the higher activities that are formed by some of these segment eventualities. This is also shown in Table 3, where a *Narration* relation that holds between the last segment of π_1 ($e_{\pi_{1b}}$) and the first segment of π_2 ($e_{\pi_{2a}}$) to constitute the overarching 'goodbye between parent and child' event labelled π_0. Thus, in the SFDRS used in this study, higher grain activities that form a narrative event are described at the highest grain level of discourse, the segment eventualities that form the higher activities are described at the mid-grain level, and the multimodal interactions used to infer segment eventualities are described at the lowest grain level.

3.3 Comparative Analysis

After the narrative events of the three film extracts have been transferred to SFDRS, the three derived representations can be used for precise comparison. Firstly, referents in the formula of defeasible eventualities of the same and (in limited cases) associated segment eventualities belonging to the same higher activities are compared for similarities. Secondly, segment eventuality types and their structure (order and discourse relations holding between similar segment eventuality types) within higher activity types are compared for similarities. And lastly, higher activity types and their structure (order and discourse relations holding between similar higher activity types) within their respective 'goodbye between parent/guardian and child' events are compared for similarities. Reoccurring configurations have been recorded during each comparison and the results compared between every two film extracts and then between all three film extracts, to find out how many configurations were repeated in two and then in all three films.

For the purposes of this pilot study, conventions were deemed to be repeated configurations in events at the levels of discourse specified above, including similar higher activity types, segment eventuality types, composition used to infer segment eventualities, and discourse relations. Conventions were sought in both single and multiple phenomena occurring simultaneously. Discourse relations patterns were not considered conventions unless found in a continuous sequence or occurring simultaneously with a convention in a segment eventuality or the composition used to infer its eventuality. The reason for this is that in many cases the same discourse relations might be expected to occur simply by virtue of the structure of film as a narrative medium as such. These relations then form the structure in every film. One example of a similar pattern found spread across a segment eventuality type, composition of segment eventuality, and discourse structure that can be found in the comparison of *Old Boy* and *Sympathy for Lady Vengeance* is illustrated in Figure 1.

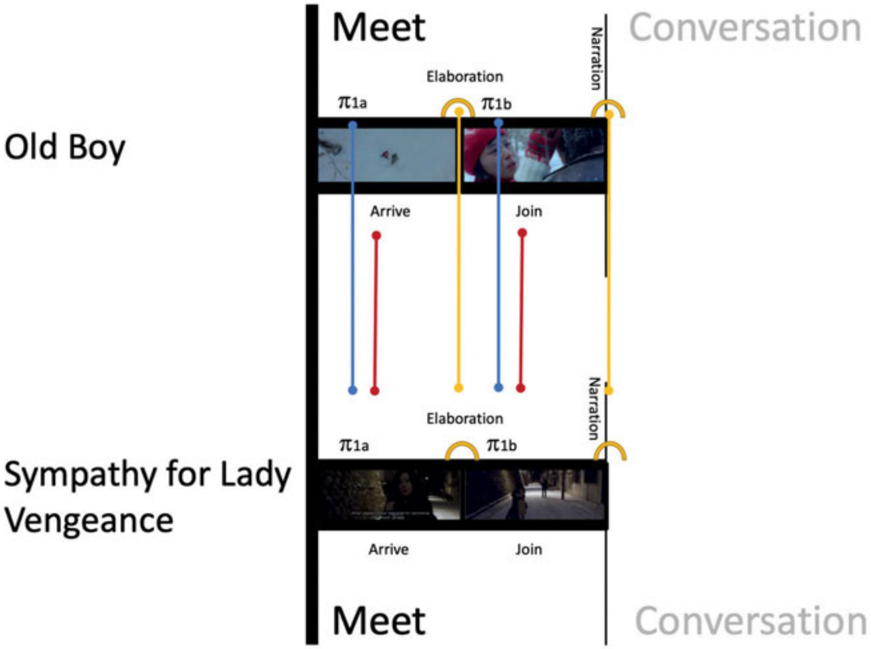


Fig. 1: Graphical representation of higher activities Meet in *Old Boy* and *Sympathy for Lady Vengeance*. We use color coding to demonstrate the patterns found through the process of comparing SFDRS. Similarities found in comparison of eventuality types that form the higher activity (red), the formulae of their defeasible eventualities (blue), and the discourse relation structure of segments (yellow) can be seen here. Further, similarity in higher activities and their order and in the order of segment eventualities is also visible according to the labelling.

4 Data Description

In this section, the SFDRS of the 'goodbye between parent/guardian and child' events from *Old Boy*, *Sympathy for Lady Vengeance*, and *The Man from Nowhere* will be presented graphically in diagrams developed by this study to accommodate higher activities (see Figures 2 and 3). Due to the number of segments in each extract and the number of resources described within each logical form, the description of the compositions used to infer segment eventualities will only be considered when appropriate in the results section below.

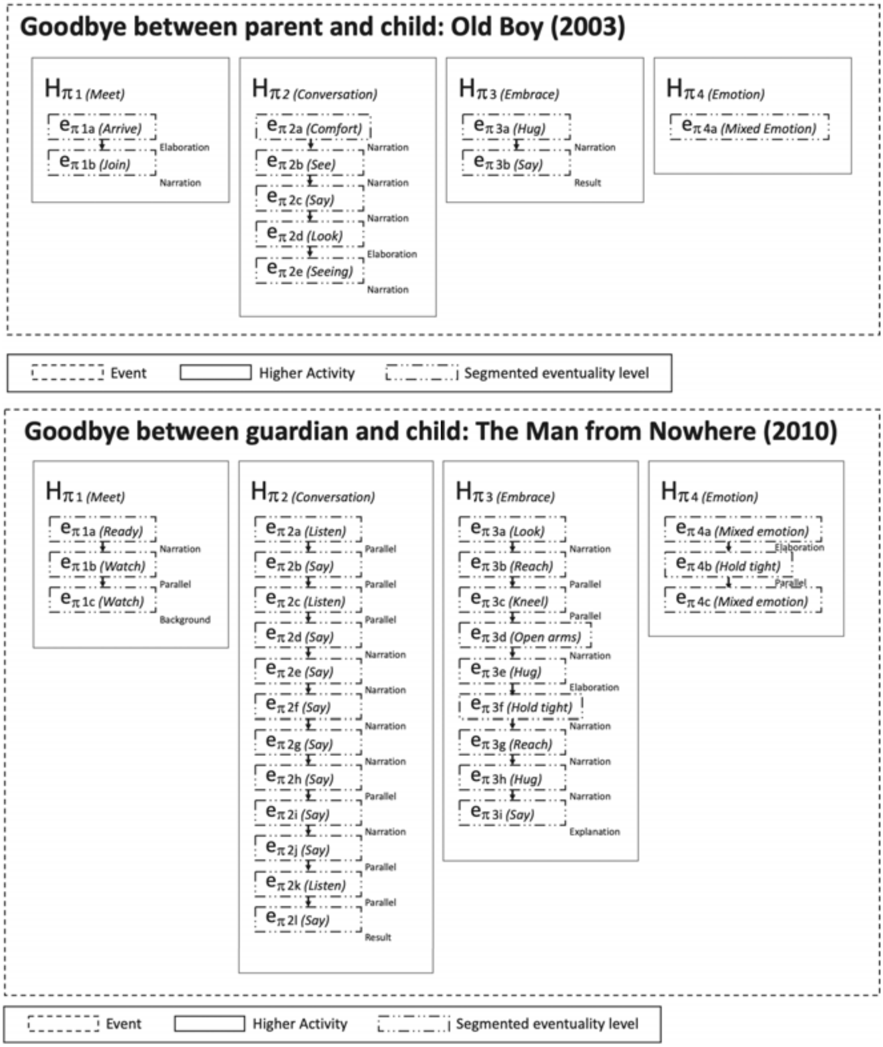


Fig. 2: A graphical representation of the SFDRS of the 'goodbye between parent/guardian and child' events extracted from *Old Boy* (upper) and *The Man from Nowhere* (lower).

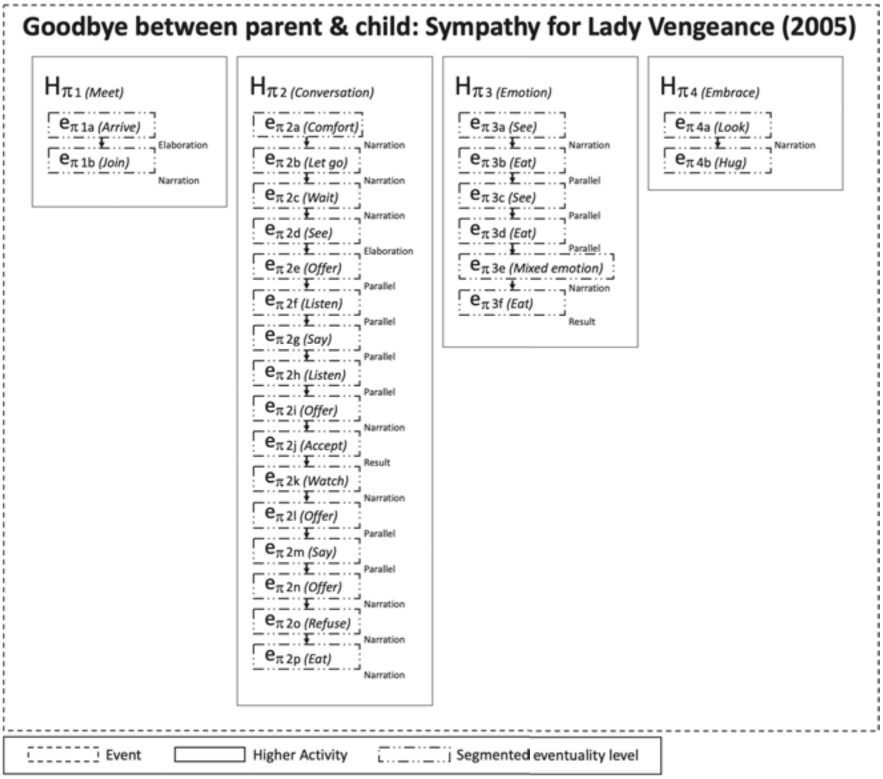


Fig. 3: A graphical representation of the SFDRS of the ‘goodbye between parent and child’ event extracted from *Sympathy for Lady Vengeance*.

5 Results

As can be seen in Figures 2 and 3, similar structures and content were found in all three film extracts at high, mid, and fine-grain levels of discourse. The most significant pattern found in all three film extracts is visibly apparent in the structure of information at the higher level of discourse. The types and quantity of higher activity were the same: Meet, Conversation, Embrace, and Emotion. There are no higher activity types in any of the SFDRS that do not exist in all three events. The order is also similar. There is only one difference in the order of the two final higher activities in *Sympathy for Lady Vengeance*, in which Emotion is followed by Embrace instead of Embrace followed by Emotion as is the case in *Old Boy* and *The Man from Nowhere*. However, one might argue that the order is in fact the same in *Sympathy for Lady Vengeance* since, in the scene, there is actually a continuation of emotion that extends from Emotion into Embrace. These similarities in type and order suggest a convention in the high-grain level of construction of the ‘goodbye between parent/guardian and child event’.

The contents of respective higher activity types were also found to be similar, and likewise suggest that there are conventions in the construction of respective higher activity types. This was emphasized by the lack of overlapping patterns in the content of Emotion and Embrace when compared between *Sympathy for Lady Vengeance* and either of the other two film extracts. Patterns in the contents of higher activities were found to be precisely confined within the boundaries of their own higher activity types regardless of the difference in order. For instance, the higher activities of Emotion in distinct films always shared patterns in their composition while Emotion and Embrace did not. This was consistent with what was found throughout the analysis: repeated configurations generally occur in the construction of similar types of narrative information at high, mid, and fine-grain levels of discourse. This is particularly noteworthy given the fact that the possibilities for constructing these events filmically are actually considerably more varied.

A larger number of patterns were found at fine- and mid-grain levels of discourse when comparing one film extract to another rather than by comparing all three film extracts at once. For instance, in comparison of *Old Boy* and *Sympathy for Lady Vengeance*, segment eventualities and the configurations used to infer them were similar overall. So much so in fact, that in *Old Boy*, which has fewer segments than *Sympathy for Lady Vengeance*, only two segments did not share a pattern in segment eventuality type and, in one of these cases, a pattern was still found in the composition used to infer it due to the segments having associated content. Although there were more segments in the extract from *Sympathy for Lady Vengeance* than in the extract from *Old Boy*, higher activities Meet and Embrace had the same quantity of segments respectively. The most significant pattern was found in the higher activities Meet. These also had the same types of segment eventuality, were in the same order, and were linked by the same discourse relation structure with no divergences from the shared pattern in either extract. Furthermore, patterns found in discourse relations were structural in every case, in that they occurred simultaneously with patterns in segment eventuality type and the compositions by which they were inferred.

In comparison of *Old Boy* and *The Man from Nowhere*, patterns were found at the mid-grain level in segment eventuality types in the higher activities Conversation, Embrace, and Emotion. The order of segment eventualities also shared a pattern in Embrace. It is, however, important to take account of the difference in the size of the segmentation in *The Man from Nowhere* compared to *Old Boy*, which inevitably resulted in larger sections of divergence from patterns in *The Man from*

Nowhere. The most evident patterns found were between Embrace higher activities and Emotion higher activities respectively. A structural pattern was found between Embrace higher activities in segment eventuality types and in the composition used to infer them, and the discourse relations holding among them. This pattern spanned this entire higher activity in *Old Boy* without a single divergence. Similarly, in the Emotion higher activity in *Old Boy* there was a pattern in segment eventuality types and the composition used for their inference that spanned the entire higher activity — although, in this case, without an accompanying pattern in discourse relations. In comparison of *Sympathy for Lady Vengeance* and *The Man from Nowhere*, patterns were found at the mid-grain level both in segment eventuality types in Conversation, Embrace, and Emotion, and in their order. The most evident pattern found was in the higher activity Conversation, in which structural patterns were found that included patterns in segment eventuality types, the compositions used to infer them, and the discourse relations holding between them. In fact, the similarities found in Conversation were so extensive that in *The Man from Nowhere* eleven out of twelve segments shared a pattern with those in *Sympathy for Lady Vengeance*. The majority of these segments had patterns at the fine-grain level of discourse, and segments had a similar discourse relation structure throughout.

6 Discussion

The number of conventions that could be identified using SFDRS appear to be dependent upon the following factors, each of which should be taken into account in future analysis:

1. **The granularity of the discourse** — At the higher level of discourse, conventions were almost entirely the same in structure and content (see Table 4). However, a larger number of differences were found at the mid-grain level of discourse in the segment eventuality types that compose higher activities, and an even larger amount again found at the fine-grain level of discourse in the composition used to infer segment eventualities
2. **Contextual similarity** — In *The Man from Nowhere* the main characters of the guardian and the child do not travel to meet each other, as the parent and child do in the other two films. Instead, they have already met and travelled together to the location where they say their final goodbye. They purchase items at a shop, prepare the child's bag, and then the camera takes the viewer outside of the shop where the two characters have grouped, and stand paused before one another. This is briefly followed by shots of the police waiting in

Tab. 4: The convention found at the high-grain level of discourse. The structure and content were almost the same in all three film extracts.

Films	Higher activity type and order			
Old Boy	Meet	Conversation	Embrace	Emotion
Sympathy for Lady Vengeance	Meet	Conversation	Emotion	Embrace
The Man from Nowhere	Meet	Conversation	Embrace	Emotion

the background and the shopkeeper watching from a distance before the camera returns to the guardian and child once again and their final interaction commences. This back-and-forth technique clearly establishes the new surroundings of the two characters and in so doing separates the preparation for the goodbye from the goodbye itself. This results in the Meet higher activity being subtly different from Meet higher activities in the other two film extracts, and subsequently having enough associated content for a pattern to be found in the inference of the defeasible eventuality but not in the segment eventuality type. This can be seen in the figures in Section 4 in segment eventualities Join (*Old Boy* and *Sympathy for Lady Vengeance*) and the segment eventuality Ready (*The Man from Nowhere*). In contrast, in *Old Boy* and *Sympathy for Lady Vengeance* there is a repeated pattern in segment eventuality types, the compositions used to infer them, and the discourse relations throughout higher activities Meet.

Greater numbers of patterns were also found in comparisons of two rather than three film extracts, which is also potentially due to contextual similarity since the larger the number of film extracts, the greater the variation there will be between the contexts of narrative events overall. For example, a pattern was found in eventuality type in all three films once per higher activity. Also, there were a total of twelve patterns found in the composition used to infer segment eventualities of the same or associated type when comparing all three films (see Table 5). While in the comparison of just a single higher activity, Conversation, in *Sympathy for Lady Vengeance* and *The Man from Nowhere* eleven patterns were found in the composition of the same and associated segment eventuality types. Furthermore, on only one occasion did all three films share a pattern in discourse relation; this occurred in the transition from higher activity Meet into higher activity Conversation.

3. **The quantity of segmentation** — Patterns consumed a significantly larger proportion of *Old Boy*'s SFDRS and were more structural than in the other film extracts which had more segments.

Tab. 5: Total patterns identified in comparison of the composition used to infer same and assocated eventuality types in all three film extracts.

Segment eventuality type	Old Boy	Sympathy for Lady Vengeance	The Man from Nowhere
Join (1b)/Join (1b)/Ready (1a)	child stands before parent	child is revealed holding parent who kneels before them	child stands (rear view) with guardian visible partially behind her
Say (2c)/Say (2g)/Say (2e)	child speaks to parent	parent speaks to child	child speaks to parent
Hug (3a)/Hug (4b)/Hug (3h)	child hugs parent	child hugs parent	child hugs guardian
Mixed emotion (4a)/Mixed emotion (4e)/Mixed emotion (4a)	Parent cries, smiles, and has tense expression	Parent cries, smiles, and has tense expression	Guardian cries, smiles, and has tense expression

4. **Sample criteria** — Conventions between *Old Boy* and *Sympathy for Lady Vengeance*, which have the closest production criteria, were the greatest overall. Tables 6 and 7 highlight the similarity of their structures and content at high, mid, and fine-grain levels of discourse.

Tab. 6: *Old Boy* and *Sympathy for Lady Vengeance* were the only two film extracts found to have the same discourse relation structures between higher activities, despite the reverse positioning of higher activities Embrace and Emotion.

Film	Between higher activities	Discourse relation
Old Boy	Meet – Conversation	Narration
	Conversation – Embrace	Narration
	Embrace – Emotion	Result
Sympathy for Lady Vengeance	Meet – Conversation	Narration
	Conversation – Emotion	Narration
	Emotion – Embrace	Result

The SFDRS produced by this study was consistent in describing what can be considered broadly shared cultural meanings. By this we refer to meanings that people will in all likelihood be able to understand regardless of differences in language or culture. For example, ‘say’, ‘look’, ‘offer’, and ‘hug’ are all happenings which are likely to be widely understood. However, the socio-pragmatic communication

Tab. 7: In comparison of *Sympathy for Lady Vengeance* and *Old Boy*, clear parallels in structure and content were found in their composition at high, mid, and fine-grain levels of discourse. *Italics* signify patterns found in the composition used to infer segment eventualities but not in the segment eventuality type, while non-italics signify patterns found in both.

Location of segment eventuality	**Old Boy**	**Sympathy for Lady Vengeance**
Meet	Arrive (1a), Join (1b)	Arrive (1a), Join (1b)
Conversation	Comfort (2a), See (2b), Say (2c), *Seeing (2e)*	Comfort (2a), See (2d), Say (2g), *Offer (2e)*
Embrace	Hug (3a)	Hug (4b)
Emotion	Mixed emotion (4a)	Mixed emotion (3e)

that takes place through Korean language and behavior were not yet singled out in this analysis. For this the predicates used to construct SFDRS would need to be systematically extended to cover the eventualities at work. In order to utilize SFDRS to its full potential in K-film analysis, therefore, it would benefit from the development of such a systematic process accounting for Korean social reasoning and how this manifests meaningfully in Korean language and behavior. The aim of this would be to enable researchers, whether they be Korean or non-Korean, to consistently recognize and label the logical forms and interactions between Korean language elements and behavior that give rise to socio-pragmatic meanings in K-film. Subsequently, K-film studies would benefit from both the formal, comprehensive description, provided by SFDRS, and a long-awaited contextualization of the meaning systems of Korean culture and language.

7 Conclusion

The aim of this study was to evaluate whether SFDRS could be utilized to find conventions in how meaning is constructed in K-film, focusing specifically on the 'final goodbye between parent/guardian and child' event. By transferring film extracts to SFDRS and comparing the resulting representations across films, it was possible to identify certain reoccurring patterns in their configurations. We suggest that these then stand as good candidates for conventions among these films. Patterns found strictly within the boundary lines of meaning types at every level of discourse supports the validity of the patterns found being conventions in types of meaning and suggests that these conventions may well be found when similar events are considered in further films.

While empirical work requires considerable effort to execute, the results of this study demonstrate that conventions benefit from more empirically-supported testing. The advantages of an empirical approach are that in contrast to single qualitative analyses, it is possible to identify clear, detailed patterns in the construction of meaning in multiple films that can then be subjected to larger scale evaluation with respect to further films as corresponding analyses become available. Further, as we demonstrated in Section 6, insights on the factors that affect identification of conventions can be gained through empirical analyses of this kind, thereby moving us towards a deeper understanding of how conventions work in film.

Filmography

Old Boy (2003). Park, C., South Korea: Show East (KR).
Sympathy for Lady Vengeance (2005). Park, C., South Korea. CJ Entertainment.
The Man from Nowhere (2010). Lee, J., South Korea: CJ Entertainment.
Zinda (2006). Gupta, S., India: Eros Entertainment.

Bibliography

Asher, N. & A. Lascarides. 2007. Segmented Discourse Representation Theory: Dynamic Semantics with Discourse Structure. In H. Bunt & R. Muskens (eds.), *Computing Meaning 3*, 87–124. Dordrecht: Springer.

Asher, Nicholas & A. Lascarides. 2003. *Logics of Conversation*. Cambridge: Cambridge University Press.

Bateman, John A. & K.-H. Schmidt. 2012. *Multimodal Film Analysis: How Films Mean*. London: Routledge.

Bateman, John A. & J. Wildfeuer. 2014. A Multimodal Discourse Theory of Visual Narrative. *Journal of Pragmatics* 74. 180–218. https://doi.org/10.1016/j.pragma.2014.10.001.

Bentley, B.P.E. 1995. The Film as Text. *Forum for Modern Language Studies* XXXI. 1–7.

Choe, S. 2009. Love Your Enemies: Revenge and Forgiveness in Films by Park Chan-Wook (Interview). *Korean Studies* 33. 29–51.

Choi, J. 2010. *The South Korean Film Renaissance: Local Hitmakers Global Provocateurs*. Middletown: Wesleyan University Press.

Cumming, S., G. Greenberg & R. Kelly. 2017. Conventions of Viewpoint Coherence in Film. *Philosophers' Imprint* 17. 1–28.

Dal, Y. 2016. *New Korean Wave: Transnational Cultural Power in the Age of Social Media*. Illinois: University of Illinois.

Han, S.H. 2011. *Gobalization and Hybridity of Korean Cinema: Critical Analysis of Korean Blockbuster Films*: University of Texas dissertation.

Kamp, Hans. 1981. A Theory of Truth and Semantic Representation. In J. A. Groenendijk, T. Janssen & M. B. Stokhof (eds.), *Formal Methods in the Study of Language. Part 1* (Mathematical Centre Tracts 136), 277–322. Amsterdam: Mathematisch Centrum Amsterdam.

Kamp, Hans & U. Reyle. 1993. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory* (Studies in Linguistics and Philosophy, Volume 42). London, Boston and Dordrecht: Kluwer Academic Publishers.

Kiaer, J. 2019. Translating Invisibility: The Case of Korean-English Literary Translation. In J. Guest & X. Li (eds.), *Translation and Literature in East Asia: Between Visibility and Invisibility*, 81–120. Oxon: Routledge.

Kiaer, J. 2020. *Pragmatic Particles: Evidence from Asian Language*. London: Bloomsbury.

Kiaer, J., A. Yates & M. Mandersloot. 2020. *On Translating Modern Korean Poetry*. London: Routledge.

Kim, S.-K. 2006. Renaissance of Korean National Cinema as a Terrain of Negotiation and Contention between the Global and the Local: Analyzing Two Korean Blockbusters, Shiri (1999) and JSA (2000). https://api.semanticscholar.org/CorpusID:5245941 (last accessed: 1 September 2021).

Morey Hawkins, J. 2018. Textual analysis. In M. Allen (ed.), *The SAGE Encyclopedia of Communication Research Methods*, 2. London: Sage.

Ok, H. 2009. The Politics of the Korean Blockbuster. In D. Kim (ed.), *Transnationalism and Film Genres in East Asian Cinema*, 37–47. Los Angeles: University of Southern California.

Oxford English Dictionary. 2020. Coherence. https://www.lexico.com/definition/coherence (last accessed: 1 September 2021).

Smith, I.R. 2013. OldBoy Goes to Bollywood: Zinda and The Transnational Appropriation of South Korean ‘Extreme’ Cinema. In A. Peirse (ed.), *Korean Horror Cinema*, 188–189. Edinburgh: Edinburgh University Press.

Wildfeuer, Janina. 2014. *Film Discourse Interpretation. Towards a New Paradigm for Multimodal Film Analysis* Routledge Studies in Multimodality. London and New York: Routledge.

Willemen, P. 2002. Detouring Korean Cinema. *Inter-Asia Cultural Studies* 3. 167–186.

Yecies, Brian & A. Shim. 2015. *The Changing Face of Korean Cinema*. London: Routledge.

Zacks, J.M., N. K. Speer, K. M. Swallow & C. J. Maley. 2010. The Brain’s Cutting-Room Floor: Segmentation of Narrative Cinema. *Frontiers in Human Neuroscience* 4. 1–24.

Dušan Stamenković and Janina Wildfeuer

An Empirical Multimodal Approach to Open-World Video Games

A Case Study of *Grand Theft Auto V*

Abstract: This chapter presents results of an empirical case study based on annotating an open-world video game and creating datasets in order to describe the semiotic elements of this video game at work. Following the procedure of identifying elements and semiotic modes on canvases of real-time video games presented in Bateman et al. (2017b), it provides a semiotic inventory of *Grand Theft Auto V* (Rockstar North 2013) so as to allow a systematic empirical analysis of the ways in which various semiotic elements are employed in the game's main story missions. This analysis of the combinations of multimodal elements across gameplay stages shows the diversity of features that structure our experience of the game and guide us within the open world. In particular, it shows mission-related and gameplay-related instructions as well as the specific result that many of these instructions last until the very end of the game. Along with several other findings, this reveals some new facets of complex mainstream game design.

Keywords: video games, semiotic elements, canvas, instructions, gameplay

1 Introduction and Motivation

Video games as dynamic, interactive audio-visual media are one of the most complex multimodal artifacts with constantly evolving new technologies and designs. Combining an immense range of semiotic elements such as moving images, colors, music, language, animation, interactive control elements, etc., various patterns of meaning construction are continually and simultaneously at work on several levels. A challenging question for the empirical analysis of these artifacts is thus the comprehensive identification of all these elements, patterns, and levels both within individual video games as well as within a specific game genre.

This chapter takes this challenge of addressing the immense range of meaning-making elements as its starting point and aims at a systematic and empirically-grounded analysis of the semiotic elements of *Grand Theft Auto V* (Rockstar North 2013) in its main story missions. Our focus explicitly lies on all kinds of semiotic elements available in the game. Here, we deliberately refrain from defining and analyzing these elements as semiotic modes, since the identification of semiotic

https://doi.org/10.1515/9783110725001-011

modes in general, and in video games specifically, will need further empirical examination on the basis of existing theoretical foundations and systematic definitions (cf. e.g., Bateman 2016).

The *Grand Theft Auto* video game series started in 1997 when events of the game's story were presented in a simple top-down perspective. From 2001 on, a far more realistic third-person perspective (sometimes shifting to a first-person view) became its standard. Over two decades of development, different technological changes and possibilities modified not only the game's appearance, but also its mechanics and overall gameplay, while keeping its popularity.

In our current discussion, we will focus on the series' latest instalment, *Grand Theft Auto V*, in which a player takes on the role of a criminal conducting different jobs for various contractors. The game is an action-adventure game developed by Rockstar North and released by Rockstar Games in 2013. Players can switch between three main protagonists (Michael de Santa, Trevor Philips, and Franklin Clinton) and freely explore an open world set within the fictional state of San Andreas, which is very similar to the Southern California region. Although the open gameworld lets players roam freely, the game's and its story's progress depend on completing various missions as well as a set of side activities. These missions are linear scenarios with predefined objectives and they are allocated to individual characters of the story in a certain order. Some missions can therefore only be unlocked after certain other missions are completed.

The game is dynamically designed and often driven by regular shootings. Side activities such as scuba diving, tennis, darts, golf, triathlon, and BASE jumping[1] all have their own rules and structures and are multimodally specific, e.g., through the use of specific colors, sounds, scenery, etc. Game players can use melee attacks, various firearms and different types of explosives to fight enemies, and can run, jump, swim, or use vehicles to navigate the world. The range of vehicles a player can control is very broad and it starts with cars and motorcycles, includes different types of vans, industrial, emergency and military vehicles, and ends with boats, helicopters, and airplanes. There are also heist missions in which the player will need skilled AI-controlled accomplices, providing help in tasks such as computer hacking and driving. One prominent interface element are 'wanted stars' displayed on a meter which appears in the upper right corner of the main canvas and shows the current wanted level, which, in turn, indicates the strength of the police force that chases the controlled protagonist (ranging from zero to five). There is also an extensive list of the so-called collectibles in the game, which includes spaceship

[1] The acronym BASE stands for building, antenna, span, and earth — objects from which one can jump.

and submarine parts, stunt jump locations, hidden packages, letter scraps, etc. All this adds to the overall complexity of the game.

With our focus on a single game, the analysis presented in this chapter should be understood as an illustrative case study that gives insights into the diversity and complexity of one typical instance from the genre of action-adventure open-world video games. Like other video games belonging to this and related genres, *GTA V* combines two different ways of presenting the player with a challenge — emergence and progression, as defined by Jesper Juul (2002). The former refers to simple rules combining and leading to variation, whereas the latter involves serially introduced challenges. The main story missions are related to different job-givers, and although there is a limited freedom in choosing the order in which one will take various jobs, there is an overall stability of the story line.

Within its genre, *Grand Theft Auto V* stands out as the third-best-selling video game of all time with over 120 million copies sold worldwide, according to Tassi (2020). One reason for this could be the immensely rich open-world design which invokes a degree of ergodic[2] work by the players, partly stirred by the depiction of the real world in the game (see Rambush & Susi 2008). This depiction contains aspects that (at least) temporarily give a linear or limiting structure to the world's freedom by means of both visual and auditory elements that appear over the usual gameworld contents.

With the superimposed aim of identifying and categorizing all these elements systematically, the first result of our study is a *semiotic inventory* of *GTA V*. In a second step, this inventory allows us to analyze (1) frequencies of various elements in the gameplay missions of this game, (2) patterns of their occurrence, as well as (3) correlations among them (i.e., their co-occurrence). With these, we are then, on the one hand, able to consider in further detail how the semiotic elements occur across the course of the game or, more specifically, in the dynamic unfolding of the various missions. On the other hand, these calculations build a basis for interpreting specific functions of the elements and their correlations, i.e., instructing the player, maintaining the game flow, checking if the annotated elements are likely to co-occur, etc. This reveals insights in how the open world of the game is constrained for the sake of providing a structured and ludic notion of progress.

[2] With 'ergodic' the authors refer to the work of Espen Aarseth, who describes situations in and with media, in which a certain commitment is demanded from the readers, viewers or players, etc., which then 'expands' the respective medium accordingly and thus also influences the type of communicative situation. This means that players in a video game take on the role of extracting meaning from the game or through the game, thereby changing the organization or content of the game (see also Aarseth 1997).

The wide scope of our analysis builds on well-developed and robust analytical frameworks evolving from the cooperation of game studies and multimodal semiotics. We explain the theoretical background for such frameworks in further detail in the following section and describe the resulting method and analytical framework in Section 3. On this basis, in Section 4, we identify game missions for which we extract and explore the semiotic elements and their combinations. By tracking the ways in which different semiotic elements are combined throughout the main story missions, the results reveal the complexity of the seemingly smooth gameplay and present different ways in which different elements can guide the players and provide a structured notion of progress in an open world.

Overall, the aimed inventory captures the particular richness of the *GTA V* world on various levels of description that demonstrate a systematic way of approaching video games within this particular genre. Furthermore, this case study may set an example for larger examinations of video games within the same genre and across different genres.

2 Literature Review and Theoretical Background

Our study is situated within the interdisciplinary context of game studies on the one hand and multimodal video game analysis on the other. While the latter is still a relatively new field and systematic and empirical analyses are just beginning to establish themselves, the former has experienced more substantial developments mostly as part of the broader field of cultural studies, but also, for example, in more focused studies of video game discourse (cf. e.g., Paul 2012; Aarseth 2014; Bell et al. 2014; Gee 2014; Ensslin 2015; Toh 2015; Ensslin & Balteiro 2019). Multimodal approaches to video games have so far been quite rare, and at the same time quite different from one another (see Hawreliak 2018; Toh 2018).

Empirical analyses have played a rather minor role within this broad field and larger corpus-based studies are still scarce. As we outline in Stamenković et al. (2017: 13), an empirically oriented analysis of games should engage with two of their key components, (1) the game structure and game mechanics, as well as (2) the gameworld and game semiotics, as described by Espen Aarseth, who sees them as part of the empirical object in game studies (along with the gameplay as the third component). As Aarseth (2014: 488–490) points out, these two components of the game object realize the processes of gameplay in their connection with and interrelation to the player while they are actively (re-)configuring and (re-)negotiating the meanings of the game during the act of play. Whereas our former work in Stamenković et al. (2017) mainly focuses on the interaction and interpretation by the player and examines the discursive structures of the dynamically unfolding video game discourse, in this study we concentrate more on the explicit semiotic level in the analysis of the gameworld and the missions available in *GTA V*. These accounts complement each other by adding further empirical details to the overall approach of systematically analysing the processes of meaning construction in video games.

For the empirical analysis of the semiotics of the gameworld, we take the notion of *interface* as discussed in the context of digital and game studies as an important starting point (see Bolter & Grusin 2000; Bolter & Gromala 2003), mainly following the proposal by Kristine Jørgensen (2012, 2013). According to her view, the video game interface should not be limited to the heads-up display (HUD) and traditional WIMP (windows, icons, menus, pointer) features which have become part of the communicative toolset belonging to the realm of video gaming. Instead, the interface also includes other signifying elements found in the gameworld itself, such as color schemes, animations, specific objects, drawings, and different interactive elements that facilitate ludic activities. This expanded view of interfaces positions gameworlds as information systems of multimodal signification which are experienced and interpreted by the players in relation to the broader context of the ludic activity. In discussing the roles of emergence and progression, Juul (2002) notes that even in an open rule-based system, some events can still be determined, and this can be a property of the game system — some games will lead to a set of conclusions, no matter what is done, but there might be variations in these conclusions at the same time. Still, in *GTA V*, the players do actively delimit gameplay trajectories in the portrayed world interface.

Jørgensen's description of gameworlds as virtual environments includes several layers of signification, dubbed *iconic*, *emphasized*, *integrated*, *overlaid*, and *metaphorical*, each of which needs to be tackled in an adequate gameworld analysis. Such an approach allows us to deal more easily with the usual demands a game interface brings with it (cf. Bogost 2015: 72). Elements that can be overlaid onto the gameworld and that can become part of the interface or the interaction between the game and the player include auditory, visual, textual, spatial, and haptic elements, each of which has its own specificities when it comes to signification. Jørgensen's description is therefore an important point of contact for multimodal analyses, as we show in the following.

A somewhat similar, but also broader level of description is taken by Bateman et al. (2017b), who see video games as multimodal artifacts that are generally embedded into larger communicative situations. In their classification of multimodal artifacts and performances (see Bateman et al. 2017b: chap. 3.3.2), they classify video games as mutable and ergodic.

To analyze different communicative situations, including video games, in detail, Bateman et al. propose seeing the different dimensions of such situations as *canvases* carrying meaningful regularities (Bateman et al. 2017b: 101). A canvas is a site of semiotic activity, i.e., the place where meaning is constructed, for example through the interaction of spatially and temporally arranged units on the screen, through the interaction of these units with a player, or in the oral interaction of a player with another player. Both the screen and the respective interaction structures can be regarded as canvases on which and for which the different semiotic elements are then examined. By choosing different canvases and sub-canvases for the analysis, the complexity of the communicative situation, which may also include other communicative situations, can be systematically addressed. In the following section, we show how our own approach adopts this systematic analysis for the empirical study of *GTA V*.

3 Method, Data, and Framework

Bringing together the notions and concepts introduced in the previous section, the methodological approach employed in the analysis of *GTA V* mainly aims at analysing the game as a form of communication by pulling apart the communicative situations involved in playing this game.

As explained above, an important part of *GTA V*'s gameplay is processed in its main story missions which build the starting point of our analysis. These missions can typically be started by moving the controlled character into an area designated by a letter on a mini-map placed in the bottom left corner of the main canvas, which we will refer to as the map. The missions are ended by the mission passed/failed screen and with this, they represent linear elements of the overall game that can be captured quite systematically. In total, there are 80 missions that comprise the main storyline of the game. Whereas we can hardly expect that two game players will conduct any mission identically, the semiotic possibilities that get exploited in these missions from the beginning to the end are similar in each instance of gameplay. This allows us to identify the elements occurring in the missions in a systematic and consistent way.

In our study, we focus entirely on these missions and their processing in our own gameplay as well as that of skilled players whose recorded gameplay can be found on YouTube.[3] Our corpus for the study thus gathers information concerning each of the 80 main story missions available in the game. We transcoded these missions and annotated their specific elements in three stages: (1) the first part of data description happened during the period in which we ourselves played the game, (2) the second part was done while we were watching the recording of our gameplay sessions, while (3) the third stage included watching the videos of other players who played the same story missions. Therefore, our mission annotations are based on two instances of gameplay of each mission, while our own instance was 'watched' twice. Each mission and the elements of each mission were described in an Excel table by two raters (the authors) who worked together in the whole process of data collection and annotation. This close collaboration is the reason why we forego the calculation of inter-rater reliability at this stage. The descriptions from all three stages were compared and on this basis, decisions for the final entries were made. Each mission is consequently presented in our coding chart (see below) as a table row containing the annotated elements, where the options for the elements were sometimes binary and sometimes had more than two possibilities.

For each mission, we then followed the general approach by Bateman et al. (2017b: chap. 7) and decomposed the communicative situation of playing a *GTA V* mission into several canvases. This analysis involves the following sequential steps, taken directly from Bateman et al. (2017b: 230) and adjusted for our current purposes:

- identify canvases within video game screens as the main building blocks of communication with the player;
- identify the semiotic elements operating in and on these canvases;
- classify those elements into categories delineated as firmly as possible;
- build an inventory of semiotic elements.

The results of this analysis and the inventory are reported in detail in Section 4.1. On the basis of the resulting inventory, we then conducted a statistical data analysis directed towards investigating:

- the *frequencies of different semiotic elements* in the annotated missions;
- the *patterns of occurrence of these elements;*
- the *correlations existing among these elements*.

[3] See channels and users such as *GTA Series Videos* (https://www.youtube.com/channel/UCuWcjpKbIDAbZfHoru1toFg), *ThirstyHyena* (https://www.youtube.com/user/ThirstyHyena), or *Willzyyy* (https://www.youtube.com/user/Willzyyy)

With these calculations, we sought to better evaluate the importance of each annotated element and see which elements in which combinations regularly influence the gameplay and guide and constrain the players' progress. As noted above, the various elements all have their own particularities in the process of meaning construction. In order to be able to describe this meaning construction in more detail in a later step of the analysis, a first step is to see when, where, and in which combinations the elements occur.

We first checked the total number of occurrences of the tracked elements within all main story missions of the game (i.e., checked in how many out of 80 missions each annotated element appears). Since these missions are subsequently related to each other sequentially, we then checked, in a second step, the pattern of occurrence of each of the elements along the progression of the main story missions, i.e., from the first mission (1) to the last mission (80) and until we covered all missions that comprise the main storyline. The pattern of occurrence along the storyline reveals in how many and which of the analyzed missions each element appears (or does not appear) and thus gives insight into the regularities of use of certain elements. This is also an important aspect for any subsequent consideration of elements in terms of semiotic modes — see, for example, Bateman (2016) — and again lets us say more about decisions of the game-makers to provide instructions and establish the game dynamics.

Finally, in the third step, we examined whether there were associations among the occurrences of the tracked elements. We did this by comparing their respective occurrences across all missions and by measuring the associations for all tracked elements pairwise. Thus, each tracked element was treated as a binary variable, taking values (Y/N) for each of the 80 missions. To measure their associations, we took the list of 80 occurrence observations for each element and the list of 80 occurrence observations for another element and calculated the corresponding Pearson correlation value between the two lists, repeating this for all pairs of tracked elements. The Pearson correlation was selected for comparing the tracked element occurrences because, when used with binary data, it is equivalent to the Phi coefficient (or a mean square contingency coefficient), which is a measure of association between two binary variables. We then calculated statistical significance values for each association measured, correcting for multiple significance testing using Holm's (1979) method. The results of this analysis and the evaluation are reported and discussed in detail in Section 4.2. By these means, we hoped to see in more detail which elements are usually combined in missions and how their combination might influence the experience and interpretation of the game. While these relationships also need to be analyzed qualitatively, for example with regard to the discourse relations holding between them, our focus here was on the quantitative calculation of their combinations.

4 Results

We first report the results of our multimodal analysis in Section 4.1, while the results of the statistical evaluation are given in Section 4.2. The coding chart in Table 4 serves as a partial overview of our database in which the results of the first part of the analysis are represented and with which we then conducted the statistical analysis and evaluation in the second part. Due to reasons of space, the table shows the coding for the first 15 missions of all 80 annotated and coded missions. As stated in Section 3, the final entries, i.e., rows with combinations of annotated elements for all 80 missions, are based on a three-stage annotation procedure. The entire table has 80 rows, one for each mission type, and 12 columns, one for each tracked element. Some of the semiotic elements we identified in these and the other missions were binary (i.e., some of them were present within one mission, and some are absent, which is noted by a simple yes/no opposition). In some cases, however, there are additional possibilities that required a more complex annotation (e.g., different types of cutscenes, different ways of phone usage, and different kinds of off-phone messaging). Our treatment of these when calculating the associations measures is described below.

As the first row of Table 4 shows, we identified several canvases which we will describe and discuss in further detail and with regard to the elements they contain in the following section.

4.1 Multimodal Analysis of Canvases and Elements

Following Bateman et al.'s (2017b) approach to the selection of potentially relevant canvases in a communicative situation, we identified, in the first step of our multimodal analysis, for all 80 missions the following canvases: the main canvas/gameworld canvas; the map canvas; the mobile phone canvas (which appears when the phone is used for calls or short text messages); the canvas of off-phone messages (on the main canvas, sometimes directly imposed on the main screen); and the sound canvas. All of these canvases are "space[s] of possibilities for perception and action" or slices through the overall space of the mission as the communicative situation (Bateman et al. 2017b: 214). Since all canvases other than the main canvas are available only in relation to the main canvas, they are basically subcanvases to this main canvas.

Figure 1 shows a screenshot of the main canvas (0) with one of the main characters in the center. Also visible are the map canvas in the bottom left (1) and the phone canvas (2), that can be turned on or off in the bottom right corner.

Tab. 1: The first 15 missions in the coding chart for *Grand Theft Auto V*.

No	Mission	Main (World) Canvas					Map Canvas		Phone Canvas	Off-Phn. Messag.	Sound		
		Cutscenes	Hghl. Areas	Txt-mis.	Txt-gmp	Main Cnv Switch	Paths	Lndm.	Used?		Vrb. Dir.	Ndg. Cues	Phone
1	Prologue	Yes (ext)	No	Yes	Yes	-	Yes	Yes	Yes (Det o.)	No	Yes	Yes	No
2	Franklin and Lamar	Yes (ext)	No	Yes	Yes	-	Yes	Yes	No	No	Yes	Yes	Yes
3	Repossession	Yes (sh)	Yes	Yes	Yes	-	Yes	Yes	Yes	No	Yes	No	Yes
4	Complications	Yes (sh)	Yes	Yes	Yes	-	Yes	Yes	Yes	Text Msg.	Yes	Yes	Yes (+SM)
5	Chop	Yes (sh)	Yes	Yes	Yes	-	Yes	Yes	No	No	Yes	Yes	Yes
6	The Long Stretch	Yes (ext)	Yes	Yes	Yes	-	Yes	Yes	No	No	Yes	Yes	No
7	Father/Son	Yes (ext)	No	Yes	Yes	-	Yes	Yes	Yes	Lifeinv. Notif.	Yes	Yes	Yes
8	Marriage Counseling	Yes (sh)	No	Yes	Yes	-	Yes	Yes	Yes	No	Yes	No	Yes
9	Daddy's Little Girl	Yes (ext)	No	Yes	Yes	-	Yes	Yes	No	No	Yes	Yes	No
10	Friend Request	Yes (sh)	No	Yes	Yes	Comp. Deskt.	Yes	Yes	Yes	No	Yes	No	Yes (+Det)
11	Casing the Jewel Store	Yes (sh)	Yes	Yes	Yes	-	Yes	Yes	No	No	Yes	Yes	No
12	Carbine Rifles (Loud Appr.)	No	Yes	Yes	Yes	-	Yes	Yes	Yes	Text Msg.	No	Yes	Yes (+SM)
13	Bugstars Equipment (Smart Appr.)	No	Yes	Yes	Yes	-	Yes	Yes	Yes	Text Msg.	No	Yes	Yes (+SM)
14	BZ Gas Grenades (Smart Appr.)	No	No	Yes	Yes	-	Yes	Yes	Yes	No	Yes	Yes	Yes
15	The Jewel Store Job (Both Appr.)	Yes (ext)	Yes	Yes	Yes	-	Yes	Yes	No	No	Yes	Yes	No

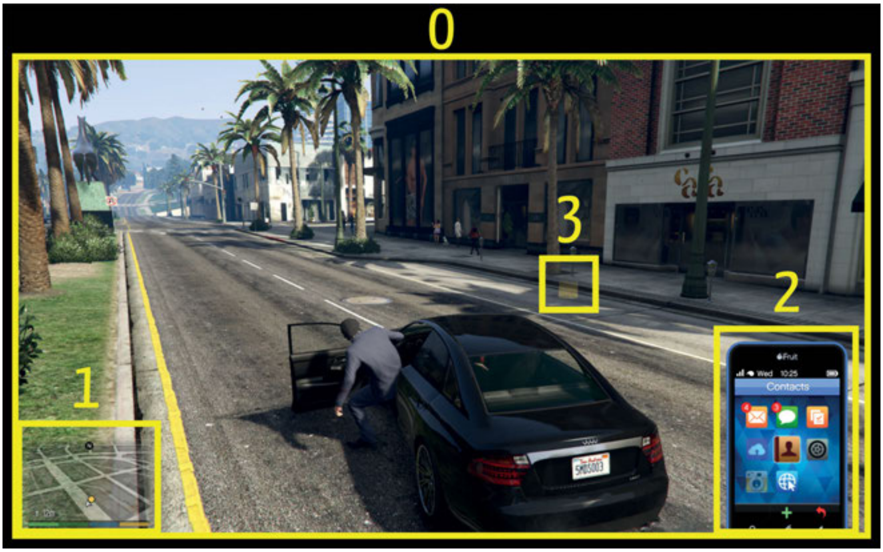


Fig. 1: *Grand Theft Auto V*'s main canvas (0), Map canvas (1), phone canvas (2) and highlighted areas (3) in the gameworld. © 2013 Rockstar Games

Furthermore, an example of a highlighted area in the gameworld is given (3). For each of these canvases, we identified the typical semiotic elements used on and in these canvases. We present both the canvases and their specific elements in the following list:

- **the main canvas/gameworld canvas**:

  cutscenes (short or extended);
  highlighted areas;
  mission-related text (including color-coded elements);
  gameplay-related text (including graphs, mostly referring to commands on the device one uses to control the game).

  All of these elements affect the gameplay. Furthermore, the main canvas has the ability to switch to more specific gameworld environments such as a computer desktop screen, phone camera, television, traffic control system, etc.
- **the map canvas**:

  usual map urban topography;
  gameplay-affecting elements: paths and designated landmarks.
- **the mobile phone canvas**:

  symbols and written text for calls and text messages.
- **off-phone messages on the main canvas**:

  written text for off-phone messages.

- **the sound canvas**:
  diegetic sounds coming from the depicted environment, including spoken verbal directions;
  non-diegetic cues (beeps and clicks) and
  phone sounds.

The map and the mobile phone can also be seen as belonging to the game's user interface, but seem more prominent and complex than elements such as the notification of wanted stars, ammunition information, weapon and character switches, so we decided to analyze them more closely. As highlighted by Bateman et al. (2017b: 374), the sound canvas "makes a different contribution to the game than, for instance, allocating yet another part of the screen to a 2D subcanvas [and it] provides access to spoken language" and other auditory cues, which we see as important semiotic elements that need to be added to the inventory.

The game also contains other canvases that could be analyzed in further detail, such as a large map and a contact list, but their use within the main story missions did not seem to be frequent enough to make them part of the coded elements for this case study. As our focus in the current approach lies on the main story missions, we also excluded elements such as the alternations of the main canvas caused by different game modes (e.g., switching to the golf-playing or tennis mode) and top-down flashes (linked to the possibility of switching among characters).

With our focus on the inventory of all semiotic elements available in the game, we did not pursue any further multimodal analyses with these canvases. However, it would for example be possible to zoom in on the main canvas and analyze discourse relations between the graphical elements and the text (either mission-related or gameplay-related) or to take the mobile phone canvas into further consideration and examine the specific multimodal interaction in and of the text messages created on this canvas. We give further examples of multimodal analyses of this kind in Stamenković et al. (2017) and Wildfeuer & Stamenković (2020).

4.2 Statistical Evaluation

For the second part of our analysis, we performed, in total, three procedures to analyze the collected data. Two of these procedures were based on the frequencies of annotated elements, i.e., their occurrences in all missions and across all missions, and one was based on correlations among the annotated elements (pointing towards their co-occurrences in missions). For these analytic purposes, the annotation chart was converted into a binary table denoting the presence or absence of all annotated elements (except for those that occurred in every mission, i.e., the mission-related text and the map landmarks). Therefore, all elements marked with “Yes”, “Yes (ext)”, “Comp. Desk.” and alike were converted into a “1”, while “No” and “-” became a “0”.

For the first step in the analysis, we examined the total number of occurrences of the tracked elements within all *GTA V*’s main story missions. In this case, we mostly treated these occurrences in a binary manner: they either occurred or did not occur within one mission yielding a maximum possible count of 80 for each element. The result, as shown in Figure 2, gives the number of missions in which we see or hear a particular element on a particular canvas.

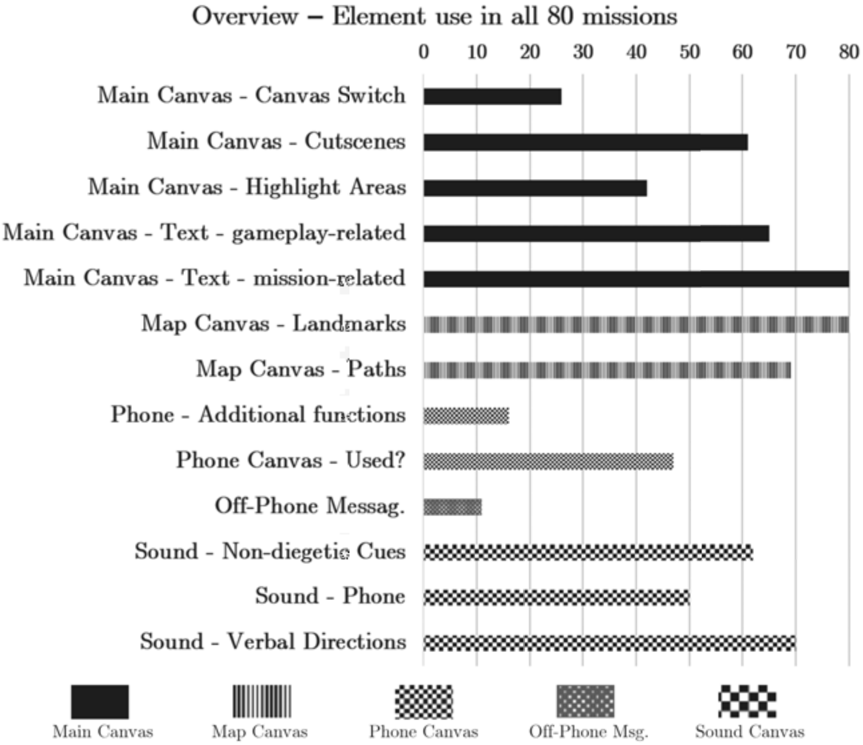


Fig. 2: The frequency of occurrence of tracked elements within all main story missions.

We can, for example, see that map landmarks (on the map canvas) and mission-related text or instructions (on the main canvas) are present in every main story mission in the game, whereas gameplay-related text (on the main canvas), map paths (on the map canvas), non-diegetic cues, and verbal directions (on the sound canvas) occur in over 60 of them. Off-phone messages and other additional functions of the phone occur less often, in under 30 of the missions, although the phone is to be heard on the sound canvas in 50 missions.

It also became visible that mission-related text co-occurs with map landmarks. Since the former frequently contains color-coded elements, there is often a cohesive link between the elements based on this color coding: for instance, if the player is instructed to “follow Lamar”, and the name “Lamar” is written in blue, there is likely to be a blue dot on the map that stands for the character that needs to be followed.

As the second step in our analysis, the pattern of occurrence of each of the elements present in Figure 2 was checked along the progression of the main story missions from the first (1) to the last one (80). Not all of them provide significant information, but some of these patterns did represent notable findings. The first of them is related to the occurrence of gameplay-related text on the main canvas screen: this text is not necessarily bound to the story, it rather provides instructions for the game player concerning *how* to play the game. It therefore includes references to controls (along with small graphs denoting particular buttons) and guides the player how to use these controls. Instances of gameplay-related text would be “Press ⟨button⟩ to toggle focus on Lamar while driving” or “Use your sniper rifle’s thermal scope to locate other snipers”. The pattern of the occurrence of gameplay-related text throughout all main story missions is given in Figure 3.

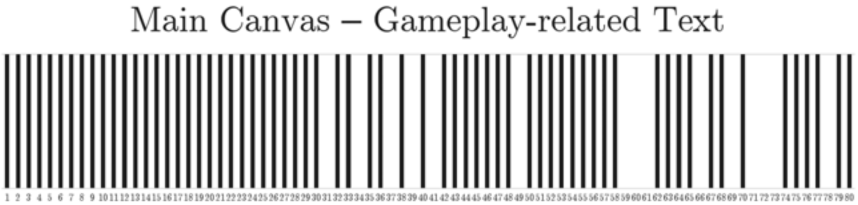


Fig. 3: The occurrence of gameplay-related text throughout all main story missions.

What we can notice here is that gameplay-related instructions appear in the initial part of the game, which is quite understandable. Although 30 consecutive missions with instructions might appear to be too many, it is reasonable to expect that the game user will be getting acquainted with the game mechanics and control within this period. However, it can also be noticed from this pattern that gameplay-related instructions actually keep appearing up to and including the very last mission.

A further result that is similarly based on tracking the occurrence of elements throughout all main story missions is related to those instances in which the whole main canvas switched to show something else. This includes a computer desktop screen, TV show, crane controller, helicopter camera, traffic control system, etc. Some of these switches were probably created to increase the feeling of being situated within an environment which resembles typical real-life situations and to improve the game dynamics. Within this category, we also included those instances in which the sniper rifle view or the car tuning view need to be used in order to complete the mission (optional uses in other missions are not counted). The pattern of occurrence of the main canvas switch is presented in Figure 4.

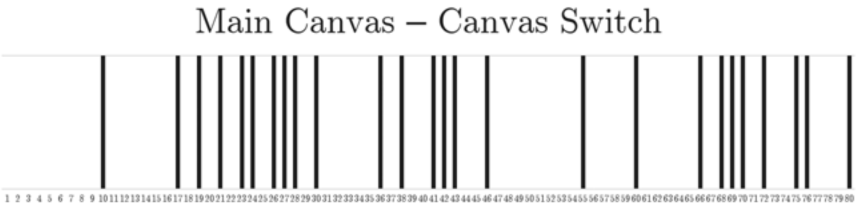


Fig. 4: The occurrence of switching the main canvas for a different one throughout all main story missions.

What is noticeable within the presented pattern is the fact that within the first 15 missions, there is only one main canvas switch, while later on they become more frequent, with sporadic stretches of missions without any main canvas switches. The reason for their absence from the initial missions is likely that this is again the period of getting accustomed to the gameplay. Occurrences become more frequent later on and it can be hypothesized that this is done in order to prevent gameplay monotony.

The pattern of this kind represents the next frequency of sounds portraying phone conversations from mission 1 to mission 80. The pattern presented in Figure 5 captures every mission where we hear at least one phone conversation of any length, regardless of whether we see the phone canvas or not (although these two elements will most frequently be correlated with each other). What becomes visible in this graph is the difference in the first and the second half of the game with regard to the number of consecutive missions in which characters communicate using the mobile phone. In the first part of the game, these instances are less frequent and less consistent, whereas in the second part there are longer stretches of missions with phone conversations. A proper interpretation of this result would definitely need to take into consideration narrative aspects of the game that explain the use of the phone with regard to the unfolding storyline or a specific character in the game.

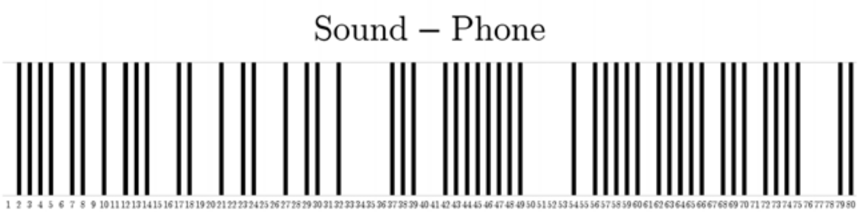


Fig. 5: The occurrence of phone conversation sounds throughout all main story missions.

The last occurrence pattern presented here is given in Figure 6. This gives us an insight related to where within the 80 analyzed missions we found cutscenes. Cutscenes are short filmic elements that happen before or after the mission gameplay itself, that precede the mission action, but also which emerge in the middle of a mission. For the purpose of achieving a consistent graph, we decided to include both the shorter and the longer cutscenes.

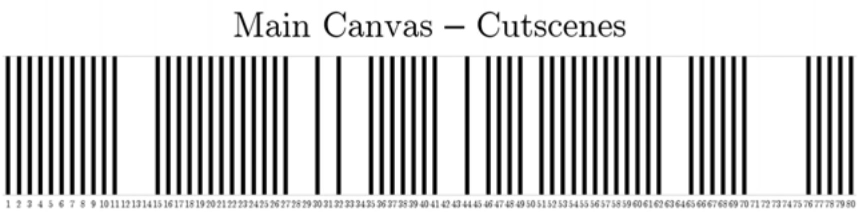


Fig. 6: The occurrence of cutscenes throughout all main story missions.

The pattern indicates that cutscenes are very frequent and tend to appear in a similar rhythm throughout the game, i.e., they are used in several sequences of missions (e.g., from mission 1 to 12, from mission 14 to mission 27, or from mission 50 to 63). However, at two points in the middle of the storyline (missions 33–34 and 42–43), we can see that the game makers left room for missions that have no cutscenes, which, as an effect, speeds up the dynamics of the game in these situations. At the beginning and in the end they are relatively constant, whereas the biggest gap can be noticed close to the game's finale (missions 71–75), when we encounter a series of vigorous missions that are not interrupted by cutscenes. The remaining occurrence tracks did not reveal patterns providing interpretable insights.

Finally, as the third step of our statistical evaluation, we present the relevant results of evaluating correlations (using Pearson's correlation coefficient as described above) existing among all elements presented in Figure 2, except for those which we found in every mission (this makes them ineligible for a correlation test). We present these results in Table 4.2.

As there have not been too many pairs with significant correlation levels after being corrected (7 in total), we are going to discuss the associations between elements that can be sensibly related to each other. For instance, there is a reason to check for correlations between the text and certain sound types, but at the same time, we cannot find a valid reason to look for correlations (or lack of correlations) between cutscenes with highlighted areas or the main canvas switch on the one hand and off-phone messages on the other.

Some of the co-occurrence patterns were expected and even obvious. For instance, the appearance of the phone canvas correlated strongly with the sound of the phone and using the phone's additional functions. However, some of them were not intuitively obvious. The strong correlation existing between seeing the gameplay-related text on the main canvas and hearing non-diegetic sounds indicates that there will be an auditory warning whenever there is such text given on the screen, which is likely to catch the player's attention. The non-diegetic sounds resemble 'incoming message' beeps one would hear on their mobile phone or laptop, but they are not generated by any of the devices we encounter in the game, which is why we have classified them as non-diegetic in the first place.

Another significant correlation is the one existing between paths on the map canvas and the highlighted areas in 'the real world' (see Figure 1), which is likely to be bound to driving to particular locations that are marked for the sake of facilitating orientation. Off-phone messaging correlates with using the phone's additional functions, which means that although in some part of the mission the textual message is presented directly on the main canvas, the phone itself can be identified as a subcanvas due to its specific elements as identified above.

5 Discussion & Conclusion

The analytical approach we used in the present study combines elements of game studies and multimodal semiotics. We applied this combination to the case study of *GTA V* as an example of an open-world video game. The data collection and coding were performed in a way that allowed an analysis that could, to a certain degree, reveal the complexity of the apparently smooth gameplay and present diverse ways in which the open world in *GTA V* can be subject to guidance and structured in order to give the player a sense of progress.

Tab. 2: The correlation chart. Note: Bold numbers indicate correlations that remained significant at the 0.01 level (2-tailed) after the *p*-values were corrected for multiple tests using Holm's method; N was 80 in all cases.

		Main Canv. Cutsc.	**Main Canv. Highl. Areas**	**Main Canv. Text-g-rel**	**Main Canv. Switch**	**Map Canv. Paths**	**Phone Canv. Used?**	**Off-Phone Msg.**	**Sound Verbal Dir.**	**Sound Non-dg. C.**	**Sound Phone**	**Phn. Add. Func.**
Main Canv. Cutscenes	Corr.	1	-.237*	.108	.074	.118	-.169	-.289**	**.411****	.051	-.068	**-.382****
	Sig.		.034	.339	.516	.296	.133	.009	**.000**	.653	.547	**.000**
Main Canv. Hghl. Areas	Corr.	-.237*	1	-.136	.019	**.420****	-.187	.089	-.132	-.093	-.220	.163
	Sig.	.034		.228	.869	**.000**	.097	.432	.242	.412	.050	.149
Main Canv. Text - g-rel.	Corr.	.108	-.136	1	-.009	.087	-.077	.099	.303**	**.892****	.025	.000
	Sig.	.339	.228		.940	.442	.496	.383	.006	**.000**	.827	1.000
Main Canv. Switch	Corr.	.074	.019	-.009	1	.122	.094	.033	.182	-.010	.152	-.013
	Sig.	.516	.869	.940		.281	.409	.772	.107	.933	.179	.906
Map Canv. Paths	Corr.	.118	**.420****	.087	.122	1	-.113	-.051	.178	.046	-.084	.109
	Sig.	.296	**.000**	.442	.281		.317	.651	.113	.688	.457	.336
Phone Canv. Used?	Corr.	-.169	-.187	-.077	.094	-.113	1	.261*	.144	-.147	**.715****	**.419****
	Sig.	.133	.097	.496	.409	.317		.019	.203	.192	**.000**	**.000**
Off-Phon. Msg.	Corr.	-.289**	.089	.099	.033	-.051	.261*	1	-.069	.128	.234*	**.526****
	Sig.	.009	.432	.383	.772	.651	.019		.545	.257	.036	**.000**
Sound Verbal Dir.	Corr.	**.411****	-.132	.303**	.182	.178	.144	-.069	1	.249*	.254*	-.094
	Sig.	**.000**	.242	.006	.107	.113	.203	.545		.026	.023	.404
Sound Non-dg.C.	Corr.	.051	-.093	**.892****	-.010	.046	-.147	.128	.249*	1	-.046	-.030
	Sig.	.653	.412	**.000**	.933	.688	.192	.257	.026		.683	.792
Sound Phone	Corr.	-.068	-.220	.025	.152	-.084	**.715****	.234*	.254*	-.046	1	.129
	Sig.	.547	.050	.827	.179	.457	**.000**	.036	.023	.683		.254
Phn. Add. Func.	Corr.	**-.382****	.163	.000	-.013	.109	**.419****	**.526****	-.094	-.030	.129	1
	Sig.	**.000**	.149	1.000	.906	.336	**.000**	**.000**	.404	.792	.254	

For instance, the results of analyzing the frequency of the various elements in the annotated missions lets us conclude that the seemingly smooth gameplay is in fact guided in a number of ways in order to provide a structured notion of progress — which appears essential in the world of mainstream video games where smoothness of this sort is desired. As pointed out in Section 4.2, several of the identified semiotic elements can be found in all or a large proportion of all main story missions (see Figure 2). In particular, the frequency of occurrence of gameplay-related text supports our hypothesis that the player is clearly and comprehensively guided through the game in spite of the open world and the range of possibilities existing in it. By being confronted with a large amount of text, the game user is exposed to a persistent 'user manual' throughout the game and is asked to learn how to play the game until the end of the main storyline (and perhaps even after that, as other missions also have their own challenges).

Various cutscenes (occurring before or within 60 missions) similarly provide scaffolding elements bound to the story and therefore also serve as a further source of guiding the player in the open world. Likewise, the use of paths on the map canvas that correlate with highlighted areas in the real world serve to facilitate orientation and to support the understanding of the gameplay. These findings most probably reflect the fact that the game design stresses the game's complexity and its plethora of possibilities. In such an open world, abundant in things one can do, the player is likely to need permanent instruction.

In addition to the frequent use of verbal-visual text and other visual elements such as map details and colored highlights, we were also able to show particular correlations of these verbal elements with auditory elements. For instance, we identified a strong correlation of gameplay-related text with non-diegetic sounds indicating a warning (see Section 4.2). This cross-modal, or intersemiotic, construction adds to the complex way of guiding the player and maintaining the multimodal game flow.

While these findings reveal some decisions of the game-makers with regard to providing instructions and establishing the game dynamics, the analysis has at the same time shown some probable limitations of our approach. In fact, the overall number of possible conclusions we could make by analyzing the data turned out to be lower than expected (especially when compared to the amount of data we had and also in relation to the time-consuming work of retrieving and manually annotating all of them). This might indicate that a more fine-grained annotation system could be applied, and/or that some elements that have been excluded should have been included (e.g., the 'obligatory' wanted stars on the main canvas would be the next candidate for addition). Another limitation lies in the need to add another analytical layer in order to be able to analyze some of the patterns of occurrence and some of the correlations with a strong link to the game's narrative.

It therefore seems promising to combine this quantitative approach to a more qualitatively oriented analysis of the story's narrative.

The idea of building a *semiotic inventory* for an individual game is clearly a strong basis for such an analysis and a combination with more abstract interpretation levels. The annotation layers for our current study could then also be imported into a larger multilevel annotation scheme, as has been suggested (mainly with regard to the page layout of comic book pages) in Bateman et al. (2017a). For now, our approach provides a substantial basis for further explorations of both individual game specificities within or across genres as well as comparative analyses of several elements in similarly and multimodally complex media. A diachronic study based on the same approach and applied to the whole *Grand Theft Auto* game series could also provide valuable results.

Bibliography

Aarseth, Espen. 2014. Ontology. In M. Wolf & B. Perron (eds.), *The Routledge Companion to Video Game Studies*, 484–492. New York: Routledge.

Aarseth, Espen J. 1997. *Cybertext: Perspectives on Ergodic Literature*. Baltimore, MD: Johns Hopkins University Press.

Bateman, John A. 2016. Methodological and Theoretical Issues for the Empirical Investigation of Multimodality. In N.-M. Klug & H. Stöckl (eds.), *Handbuch Sprache im multimodalen Kontext* (Handbooks of Linguistics and Communication Science (HSK) 7), 36–74. Berlin: De Gruyter Mouton.

Bateman, John A., F. O. Veloso, J. Wildfeuer, F. H. Cheung & N. S. Guo. 2017a. An Open Multilevel Classification Scheme for the Visual Layout of Comics and Graphic Novels: Motivation and Design. *Journal of Digital Scholarship in the Humanities* 32(3). 476–510. https://doi.org/10.1093/llc/fqw024.

Bateman, John A., J. Wildfeuer & T. Hiippala. 2017b. *Multimodality – Foundations, Research and Analysis. A Problem-Oriented Introduction*. Berlin: De Gruyter Mouton.

Bell, Alice, A. Ensslin & H. K. Rustand. 2014. From Theorizing to Analyzing Digital Fiction. In A. Bell, A. Ensslin & H. K. Rustand (eds.), *Analyzing Digital Fiction*, 3–20. New York and London: Routledge.

Bogost, Ian. 2015. *How to Talk About Videogames*. Minneapolis: University of Minnesota Press.

Bolter, Jay David & D. Gromala. 2003. *Windows and Mirrors: Interaction Design, Digital Art, and the Myth of Transparency*. Cambridge, MA: MIT Press.

Bolter, Jay David & R. Grusin. 2000. *Remediation: Understanding New Media*. Cambridge, MA: MIT Press.

Ensslin, Astrid. 2015. Discourse of Games. In T. K., C. Illie & T. Sandel (eds.), *The International Encyclopedia of Language and Social Interaction*, Hoboken, NJ: Wiley Online. https://doi.org/10.1002/9781118611463.wbielsi154.

Ensslin, Astrid & I. Balteiro (eds.). 2019. *Approaches To Videogame Discourse: Lexis, Interaction, Textuality*. New York: Bloomsbury. https://doi.org/10.5040/9781501338489.

Gee, James Paul. 2014. *Unified Discourse Analysis: Language, Reality, Virtual Worlds and Video Games*. New York: Routledge. https://doi.org/10.4324/9781315774459.

Hawreliak, Jason. 2018. *Multimodal Semiotics and Rhetoric in Videogames*. New York and London: Routledge.

Holm, Sture. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6. 65–70.

Jørgensen, Kristine. 2012. Between the Game System and the Fictional World: A Study of Computer Game Interfaces. *Games and Culture* 7. 142–163. https://doi.org/10.1177/1555412012440315.

Jørgensen, Kristine. 2013. *Gameworld Interfaces*. Cambridge, MA: The MIT Press. https://doi.org/10.7551/mitpress/9780262026864.003.0003.

Juul, Jesper. 2002. The Open and the Closed: Games of Emergence and Games of Progression. In F. Mäyrä (ed.), *Computer Games and Digital Cultures Conference Proceedings*, 323–329. Tampere: Tampere University Press.

Paul, Christopher A. 2012. *Wordplay and the Discourse of Video Games: Analyzing Words, Design, and Play*. New York and London: Routledge. https://doi.org/10.4324/9780203124031.

Rambush, Jana & T. Susi. 2008. The Challenge of Managing Affordances. *Human IT* 9. 83–109.

Rockstar North. 2013. *Grand Theft Auto V*. Rockstar Games.

Stamenković, Dušan, M. Jaćević & J. Wildfeuer. 2017. The Persuasive Aims of Metal Gear Solid: A Discourse Theoretical Approach to the Study of Argumentation in Video Games. *Discourse, Context and Media* 15. 11–23. https://doi.org/10.1016/j.dcm.2016.12.002.

Tassi, Paul. 2020. The Enduring Mystery Of How 'GTA 5' Has Sold 120 Million Copies. https://www.forbes.com/sites/paultassi/2020/02/15/the-enduring-mystery-of-how-gta-5-has-sold-120-million-copies (last accessed: 1 September 2021).

Toh, Weimin. 2015. A Multimodal Discourse Analysis of Video Games: A Ludo-Narrative Model. In S. Björk & M. Fuchs (eds.), *Proceedings of the DiGRA 2015: Diversity of Play: Games – Cultures – Identities*, Lüneburg, Germany.

Toh, Weimin. 2018. *A Multimodal Approach to Video Games and the Player Experience*. New York and London: Routledge.

Wildfeuer, Janina & D. Stamenković. 2020. Multimodale Forschungsperspektiven auf Computerspiele. In M. Engelns & P. Voßkamp (eds.), *Sprechende Pixel – Computerspielphilologie in Schule und Hochschule. (OBST 96/2020)*, 7–28. Duisburg: Universitätsverlag Rhein-Ruhr.

List of Contributors

John A. Bateman has been Professor of Applied Linguistics in the Faculty of Linguistics and Literary Sciences at the University of Bremen since 1999. His main fields of research range over computational linguistics (particularly natural language generation, discourse, and dialogue), formal ontology, and the theory and practice of multimodality. He has published widely in all of these areas, including monographs on text generation (1991, Pinter, co-authored with Christian Matthiessen), multimodality and genre (2008, Palgrave), film (2012, Routledge, with Karl-Heinrich Schmidt), text and image (2014, Routledge), and an introduction to multimodality as a whole (2017, de Gruyter, with Janina Wildfeuer and Tuomo Hiippala). Recent work focuses specifically on the semiotic foundations of multimodality and the use of empirical methods for their investigation.

Kristoffer Claussen Boesen was an undergraduate student at International Business Communication, University of Southern Denmark, at the time of this research. He currently studies software design at the IT University in Copenhagen and aims to apply those skills in the pursuit of digital humanities research in the future.

Ralph Ewerth is Professor at the Leibniz University Hannover and head of the Visual Analytics Research Group at TIB — Leibniz Information Centre for Science and Technology in Hannover, Germany; since 2016 he has been also a member of the L3S Research Center (Hannover). He received the Diploma and Ph.D. degree in Computer Science in 2002 and 2008, respectively, both from the University of Marburg, Germany. His research interests include automatic analysis of multimodal data, multimedia retrieval, and machine learning. Dr. Ewerth has published more than 80 peer-reviewed papers at international conferences and journals, and received several awards for his research.

Tuomo Hiippala is Assistant Professor of English Language and Digital Humanities at the University of Helsinki, Finland. His current research focuses on the application of computational methods to support empirical multimodality research. His major publications include The Structure of Multimodal Documents (2015, Routledge) and Multimodality: Foundations, Research and Analysis (2017, De Gruyter, with John A. Bateman and Janina Wildfeuer).

Zhanhao Jiang is Professor of Foreign Linguistics and Applied Linguistics Research Centre, Xi’an International Studies University, China. He is also a member of the Chinese Pragmatics Association. His research focuses on pragmatics, multimodality, and second language acquisition. His recent research project is on the development and application of a multimodal corpus study of students’ pragmatic competence. He has recently contributed articles to the European Early Childhood Education Research Journal.

Jiaping Kang is a PhD candidate at Xi’an International Studies University, China. Her research interests include systemic-functional linguistics, multimodality, and page-based discourse analysis. Her recent publication involves a Bibliometrix-based visualization analysis on the research focus and trend in multimodality abroad (2020). Her current research focus is on multimodal discourse and visual metaphor.

https://doi.org/10.1515/9783110725001-012

Angela Kessell (now Angela Noel) earned her Ph.D. in Cognitive Psychology from Stanford University. Her research has included gestural and spoken natural language human-computer interfaces for HRL Labs, gesture, visualization, and context-aware interfaces at Stanford, human-in-the-loop simulations for the NASA ARMD Airspace Systems program, and automatic analysis and visualization of gestural, verbal, and physiological information during human-human conversation for NASA.

Jieun Kiaer is Associate Professor in Korean Language and Linguistics at the University of Oxford, UK. She works in areas of cross-cultural, multimodal communications particularly from an East Asian perspective and has widely published in pragmatics as well as Korean linguistics and translation.

Loli Kim is a DPhil Researcher of Korean Studies at the University of Oxford, UK. She works in areas of multimodal, semantic, and cross-cultural communications, particularly from a Korean perspective. Her current research focuses on the translation of multimodal meaning-making that is ‘untranslatable’ or Anglophone European viewers of South Korean film, specifically the socio-pragmatic verbal and non-verbal behavioral expressions.

Christian Mosbæk Johannessen is Associate Professor at the Department of Language and Communication, University of Southern Denmark. His primary area of research is graphic communication from a social semiotic and eco-social perspective, and he is particularly interested in the materiality of graphics, the dynamics of graphic conventionalization, and the intersection between graphics and cognition.

Eric Müller-Budack received the Master of Engineering (M. Eng.) from the Jena University of Applied Sciences, Germany, in 2014. He is currently a Ph.D. candidate in the Visual Analytics Research Group at TIB – Leibniz Information Centre for Science and Technology in Hannover, Germany. His research interests include automatic multimedia indexing and retrieval, analysis of multimodal information, and sports analytics. In this context, he primarily focuses on deep learning approaches for multimodal news analytics to quantify cross-modal relations of entity representations in image-text pairs.

Christian Otto received the Master of Science (M. Sc.) from the Friedrich-Schiller-University in Jena, Germany, in 2014. He is currently working as a Ph.D. candidate in the Visual Analytics Research Group at TIB – Leibniz Information Centre for Science and Technology in Hannover, Germany. The focus of his research is the analysis of multimodal information. On one side, he attempts to bridge the gap between the computer science perspective and related work in media and communication sciences, and, at the same time, investigating applications to the field of search as learning in the Web and online search in general.

Jana Pflaeging is a researcher in English and applied linguistics at the Department of English and American Studies at Salzburg University. She currently pursues a binational Ph.D. at Salzburg University, Austria, and Halle-Wittenberg University, Germany. Her research interests are in multimodal genre studies and text/discourse linguistics. Trained in English linguistics and fine arts, she explores the synergies between both fields when creating visualizations of linguistic and multimodal theories, methods, and data.

Andreas Rothenhöfer is a lecturer for German linguistics at the University of Bremen. His research interests include language and emotion, discourse, and multimodal interaction.

Dušan Stamenković is an Associate Professor in the Department of English, Faculty of Philosophy, University of Niš, Chief of the Language Cognition Lab, a former Fulbright Visiting Scholar at UCLA's Reasoning Lab and an editorial team member at *Visual Communication*. His research interests include psycholinguistics, metaphor comprehension, multimodality, comics and video games studies, as well as contrastive linguistics and translation studies. He has authored one monograph, co-authored one coursebook and published over 60 papers in thematic volumes, books of proceedings, as well as journals (including *Journal of Memory and Language*, *Psychological Bulletin*, *Metaphor and Symbol* and *Discourse, Context & Media*).

Hartmut Stöckl is Full Professor of English and Applied Linguistics in the Department of English and American Studies at Salzburg University, Austria. His main research areas are in semiotics, media/text linguistics/stylistics, pragmatics, and linguistic multimodality research. He is particularly interested in the linkage of language and image in modern media, typography and an aesthetic appreciation of advertising. A volume on *Shifts towards Image-centricity in Contemporary Multimodal Practices* with Routledge (2019) is his most recent co-edited book.

Mads Lomholt Tvede was an undergraduate student at International Business Communication, University of Southern Denmark, at the time of this research. He currently studies software design at the IT University in Copenhagen and aims to apply those skills in the pursuit of digital humanities research in the future.

Barbara Tversky is a cognitive psychologist, Professor at Columbia Teachers College and emerita at Stanford. Her research includes memory, categorization, language, spatial thinking, event perception and cognition, visualization, gesture, and creativity. She has enjoyed collaborations with linguists, neuroscientists, philosophers, computer scientists, domain scientists, artists, musicians, and designers. She has served on boards of many professional associations, journals, and science outreach organizations, was president of the Association for Psychological Science and is a fellow of that association, Cognitive Science, the Society for Experimental Psychology, and the Academy of Arts and Sciences.

Janina Wildfeuer is Assistant Professor in the Department of Communication and Information Studies at the University of Groningen, NL. Her main research interests lie in the areas of multimodal linguistics, media studies, discourse analysis, and semiotics. She teaches classes in multimodal, interdisciplinary, and applied linguistics and analyses films, comics, and other multimodal documents in several projects exploring the notion of multimodal discourse. Her publications include several monographs and edited collections as well as contributions and articles on the analysis of multimodal artifacts, mostly focusing on interdisciplinary approaches in the humanities and beyond.

Index

https://doi.org/10.1515/9783110725001-013