Multidisciplinary Approaches to Discourse 2008 – MAD 08
'Linearisation and Segmentation in Discourse'
Lysebu, Oslo, 20-23 February 2008

Linearization of arguments in commentary text

Manfred Stede and Antje Sauermann

AG Angewandte Computerlinguistik

Institut für Linguistik

Universität Potsdam / Germany

stede|sauermann@ling.uni-potsdam.de

Abstract

We report on our ongoing investigation into the relationship between the linear order of text segments and the underlying *argument structure* in certain newspaper commentaries. After briefly introducing our corpus and the general layout of the research project, we describe our approach to representing argument structure as a "support graph". Then we turn to the relation between this abstract structure and the linearization of the argument in the text; to this end, we suggest a mapping between the support graph and the text linearization, and offer some first observations on correlations.

1 Research framework and corpus

Our research is embedded in the framework of *multi-level annotation*, an approach that does not aim at capturing discourse-structural phenomena in a single representation, but distributes information into several different conceptual realms and corresponding distinct technical annotation layers (see Stede 2007 and Stede 2008). Texts are analyzed on levels such as syntax, coreference, information structure, or conjunctive relations, and annotations are produced with dedicated software tools. The results of the individual annotations are stored in a database that allows for viewing the annotations, querying the data across annotation levels, and running statistical analyses to explore relationships between different levels (a step that we label "annotation mining").

The research we report here is a pilot study in which the authors carefully examined 11 texts and negotiated "gold standard" analyses of argument structure and also of rhetorical structure in line with (Mann & Thompson 1988). The experience gained in this negotiation process leads to the formulation of specific and detailed annotation guidelines. The "real" study will then involve independent annotators working solely on the basis of the guidelines. Inter-annotator agreement will be measured to check whether the task is manageable; if so, correlations between these and other annotations (on different levels) will be investigated systematically.

The corpus we use is a collection of German newspaper commentaries (Stede 2004a). For the specific research reported in this paper, we focus on a particular sub-corpus with commentaries from the *Pro & Contra* section of *Tagesspiegel am Sonntag*. These short pieces (12 to 16 sentences) reply to a yes/no question currently under debate in Berlin politics; both a "pro" and a "contra" opinion are published next to each other, accompanied by an article giving background information. Thus in these texts we find very crisp argumentation: authors have to clearly make their point and state the reasons for their opinion, thereby trying to convince readers of their position. A sample text from our collection, along with an English translation, is given in Figure 1.[1] Numbers in square brackets give our segmentation, which makes use of a few simplifications; for reasons of space, in this abstract we do not discuss our segmentation rules, which are based on (Jasinskaja et al 2007).

[1] Soll Berlin sich um Olympia 2016 bewerben?

[2] Hamburg hat es längst begriffen: [3] Olympia ist Gold wert. [4] Wer die Spiele in die Stadt holt, steht im weltweiten Wettstreit um Aufmerksamkeit auf dem Siegertreppchen. [5] Darum darf Berlin die Chance auf Oylmpia 2016 jetzt nicht verspielen. [6] Die Hauptstadt muss den Staffelstab von Leipzig übernehmen und sich als Austragungsort bewerben. [7] Barcelona hat gezeigt, dass der olympische Effekt unbezahlbar ist. [8] Mit den Spielen 1992 hat sich die Stadt neu erfunden - und macht bis heute Gewinn: [9] Die Zahl der Übernachtungen hat sich verdoppelt, die Wirtschaft profitiert noch immer. [10] Wenn sich Berlin nun im zweiten Anlauf bewirbt, zeigen wir der Welt, dass wir es besser können als einst. [11] Schließlich bringt die Stadt heute mit, was ein Kandidat braucht: [12] Metropolenflair, Hotelbetten, Infrastruktur. [13] Die für Olympia 2000 konzipierten Sportstätten wie das Velodrom und die Max-Schmeling Halle stehen, das Olympiastadion ist so gut wie neu, die Anschütz-Arena kommt. [14] Schon durch die erneute Bewerbung würde sich Berlin modernisieren und international profilieren. [15] Staatliche und private Gelder könnten fließen, Millionenzuschüsse vom IOC würden folgen. [16] Und selbst, wenn schon 2012 eine europäische Metropole Ausrichter werden sollte: [17] Man muss die Muskeln spielen lassen, um die Spiele wenn nicht mit der zweiten, dann eben mit der dritten Bewerbung in die Stadt zu holen. [18] Berlin an die Stadtblöcke: [19] Achtung, fertig, los!

[1] Should Berlin apply for the 2016 Olympics?

[2] Hamburg has long understood: [3] Olympic games are worth a lot of gold. [4] Those who draw the Olympics into their city are winners in the world-wide competition for attention. [5] That's why Berlin must not let the opportunity for the 2016 games pass. [6] The capital must grab the baton from Leipzig and apply to be the venue. [7] Barcelona has shown that the olympic effect is invaluable. [8] With the 1992 games the city has re-invented itself -- and makes profit up to today: [9] The number of overnight stays has doubled, the economy is still profiting. [10] When Berlin now runs again as applicant, we show the world that we're better now than we were once. [11] After all, today the city offers what a candidate needs: [12] big-city flair, hotel beds, infrastructure. [13] The sports venues planned for 2000, such as Velodrom and Max-Schmeling-Halle, exist, the olympic stadium is in mint condition, the Anschütz arena is nearing completion. [14] Just by re-applying, Berlin would already modernize itself and improve its international profile. [15] Public and private sponsoring money would pour in, millions would follow from IOC. [16] And even if a European city turns out to be the venue for 2012: [17] One has to flex one's muscles in order to win the games, if necessary with the third instead of the second application. [18] Berlin to the starting block: [19] On your mark, ready, go!

Figure 1: Sample text 'Olympics'

[1] To understand the text, it helps to know that some 15 years ago, Berlin had already placed an unsuccessful bid for the 2000 Olympics.

2 Representing argument

Generally speaking, identifying and formally representing the structure of argument in a text is a quite complex task (see, e.g., Reed 2006). Due to the specific nature of our *Pro & Contra* texts, however, our proposal is that a relatively straightforward representation scheme can be used to adequately capture the essence of the argument; furthermore, we hope that this scheme will lead to reliable (in the sense of inter-annotator agreement) annotations on the basis of dedicated guidelines. Our main source of inspiration is the work of Freeman (1991), which can be characterized as a "compositional" extension of the well-known argument schema by Toulmin (1958). In a nutshell, while Toulmin was interested in the nature of argument *per se*, Freeman undertook the step to isolate the components and link them to linguistic phenomena, thus enabling a piecemeal construction of specific arguments formulated in text. Still, Freeman (like Toulmin) was a philosopher – he did not go as far as analyzing "real" texts of natural length and breadth. That is why our scheme in turn extends Freeman's notation at various points.

For the most part, the argument structure is represented as a directed graph with nodes representing segments of the text and arcs showing "support" relationships between the illocutions expressed in two segments. The notion of one segment "supporting" another is the same as that proposed by Brandt & Rosengren (1992) in their *illocution structure* of texts. It can be paraphrased as: <supported claim/stmt>. *Why? Because* <supporting stmt>.

Sometimes, the support relation is explicitly marked in the text, as in [5] or [11] of our sample text; often it is only implicit and needs to be inferred. This implies that the linear order of the two elements can vary; in the sample text, [7] is a general statement which is followed, without explicit signal, by the support, namely the specific observations in [8] and [9]. Importantly, *support* is not a relation that uniquely connects two segments, nor do the segments have to be adjacent. Typically, a commentary offers a variety of reasons supporting the main claim; this leads on the one hand to nodes with multiple parents and on the other hand to parent-child links of non-adjacent text spans. Both can be observed in the analysis of our sample text in Figure 2 on the next page.[2]

While the support relationship is central to argument structure, it is of course not *sufficient* to represent it. We distinguish two further ways of linking segments, this time restricted to adjacent segments. (1) Segments are bundled together in a complex node when the second segment provides illustration or other elaboration of the first, with no recognizable support relation between the two. See [18/19] in our analysis. (2) Following Freeman, we distinguish the case where two segments *collectively* support a third one; neither the first nor the second could fulfil that function in isolation, they thus depend on each other. Often, these two are in a contrastive relation, and only the entire contrast plays the intended argumentative role. Another frequent pattern is the first sentence ending with a colon and the second serving to clarify or strengthen the point, thereby providing more than just an elaboration. An example is [8]/[9] in our analysis.

[2] In practice, we are using an annotation tool that handles nodes containing the entire text segments, which simplifies the decision process considerably. Here we show segments merely as numbered nodes just for brevity.

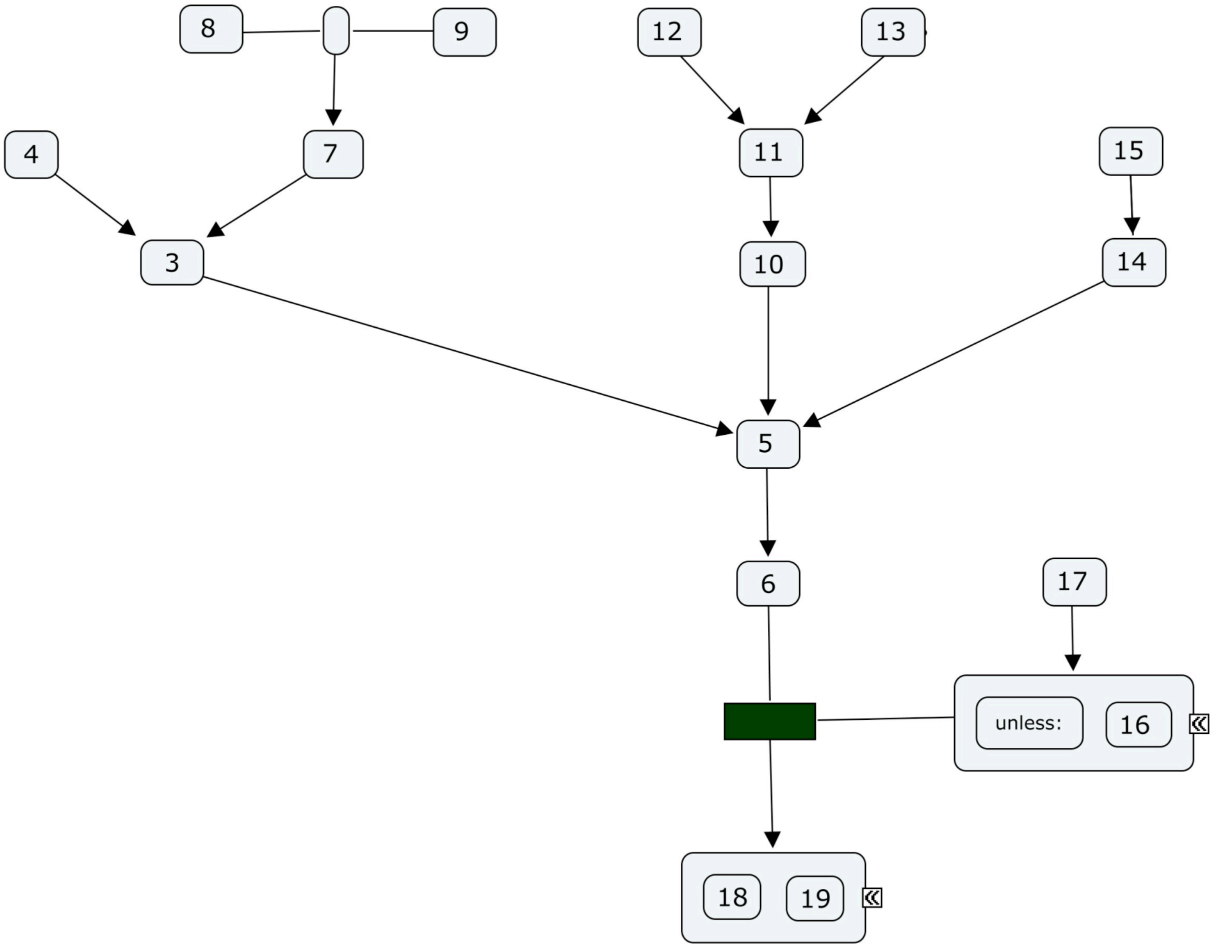

Figure 2: Argument structure of sample text 'Olympics'

One central feature we postulate for the graph is the presence of a single "root" node, i.e. a node that does not support any other. In our guidelines, we ask annotators to identify the "central" statement of the text early in the process; if successful, the corresponding node will be the root of the graph. Sometimes, however, no segment seems to clearly summarize the main message; in those cases we create an artificial root node labelled YES or NO, depending on the answer the text gives to the question stated in the headline.

Even our short, "crisp" texts can contain segments that do not play an inherent role in the argument. Instead, they serve purposes of changing topics, or they constitute rhetorical moves; one example from our corpus is an argument followed by "But we may ignore that point" ("Aber lassen wir das ruhig beiseite."), which does not at all intend to wipe out the point; instead it is a rhetorical device saying that "I have even more important arguments, which will follow." Argument structure in the sense of the "support graph" is the naked skeleton of the argumentative text – it abstracts from decorum such as statements of the kind just mentioned. In the sample text, we see [2] as decorum that introduces a topic but does not enter the support graph. Also, by convention, we always leave out the headline of the text [1], which is the question to be answered. In short, our argument structure does not necessarily completely span the text, and it can link segments that are distant from each other. (Our guidelines, however, instruct annotators to prefer "local"" links in cases of ambiguity or doubt.)

Many if not most argumentative texts make use of a highly effective rhetorical strategy: they concede a possible counter-argument to their own position and then proceed to refute it; in sum, this serves to further strengthen the own position. In the Toulmin schema as well as in Freeman's work this is labelled as *rebuttal* and *counter-rebuttal*. Following Freeman, we mark a rebuttal by a horizontal line connecting the rebuttal node to a black rectangle, which crosses the support relation(s) that is (are) "blocked" by the rebuttal.[3] As shown in Figure 2, [16] is the only rebuttal in our sample text: In case some European city is declared host of the 2012 games, Berlin's chances for 2016 will decrease. But, according to the author, this should not stop the city: the counter-rebuttal is given in [17], which is connected by a vertical line to the rebuttal node. In general, both rebuttals and counter-rebuttals can be complex themselves; if the rebuttal is supported, we build a complex node containing this sub-argument.

In short, our annotation guidelines specify the following sequence of steps: Segmentation into minimal units – eliminating units – bundling units – identifying central segment – identifying rebuttals and counter-rebuttals – identifying support relations – drawing the complete graph. The inventory of link types described above has proven sufficient to represent the structure of our 11 texts; in the next step, we will have annotators work with more than these to test the coverage.

3 Argument structure and linear order

We emphasized that our notion of argument structure abstracts a good deal from the text; we think it can in fact be regarded as a representation of the "final step" in interpreting this type of text (i.e., argumentative; as opposed to narrative, instructive, expository and descriptive – see Werlich 1975). In particular, we do not expect this structure to correlate immediately with a mental representation to be built up in *incremental* fashion (which is the focus of interest, *inter alia*, for SDRT Asher & Lascarides 2003). Instead, we regard the argument structure as the reader's reconstruction of the author's underlying plan, which requires quite a bit of interpretation work on the part of the reader, including the derivation of links between non-adjacent segments.

It follows that the relationship between argument structure and the linear order of the text need not be a very simple one. Recall that we are, for instance, deliberately allowing for multiple, possibly distant, segments supporting the same conclusion in the graph. It thus becomes interesting to investigate how any "breaks" or "hops" in the argument are coded at the linguistic surface: are readers being given explicit cues for uncovering the underlying, possibly long-distance, relationships between elements of the argument? In the framework of multi-level representation, this question is one of correlating the argument structure with other levels, in particular thematic development and sentential information structure. As a prerequisite, however, we need to adequately map the linearization-relevant information from the support graph to the individual segments.

For this purpose, we have devised a set of labels that for a pair of segments adjacent in the text, characterize the "topological" relation holding in the support graph. That is, we proceed in linear fashion from segment to segment in the text and record the corresponding relation between the two segments in the support graph. For the example text and its argument structure (cf. Figures 1 and 2), the labels of the segments are as follows:

[3] In graph-theoretic terms and in the underlying XML representation, the black rectangle corresponds to an artificial extra node.

> 1: DEL / 2: DEL / 3: NEW / 4: REASON / 5: CONC* / 6: JOIN / 7: REASONp / 8: REASON / 9: LINK / 10: REASONp / 11: REASON / 12: REASON / 13: SISTER / 14: REASONp / 15: REASON / 16: NEW-REB / 17: COU-REB / 18: CONCp / 19: JOIN

For 1 and 2, labels DEL(ete) indicate that the segments do not show up in the support graph at all (because they do not contribute to the underlying argument). For all other segments, the label marks the relation holding between the segment and its predecessor in the graph. NEW (3, 16) states that the segment "jumps" to a new node in a yet unknown path in the graph; this node can be part of the rebuttal box, in which case the label is NEW-REB (16). In "simple" cases, a segment is a REASON (4, 8, 11, 12, 15) or a CONC(lusion) of the preceding segment – i.e., adjacency in the text corresponds to adjacency in the support graph. When we stay on the same path in the graph but do not move to an immediate neighbour, we add one or more asterisks to the label; thus 5 moves to a conclusion of a conclusion of 4. Often, we move back to a subgraph that was already "seen", which is indicated by a 'p' at the end of the label: REASONp (7, 10, 14) and CONCp (18) state that the segment's node is a reason or a conclusion of a node we have already visited. The two ways of bundling adjacent segments are marked with JOIN (6) and LINK (9), respectively. SISTER (13) indicates that the segment supports the same conclusion as its predecessor – we open a new path in the immediate neighbourhood, so to speak. Finally, COU-REB (17) marks a move from the rebuttal subgraph to the counter-rebuttal subgraph.

As shown above, the sequence of labels arises by following segments in linear order and checking the kind of relationship present in the graph. It obviously does not contain *all* information from the support graph but deliberately selects only the information that pertains to linearization. The labels now allow us to readily identify those segments that do not simply "continue" the argument flow from the previous segment but represent a "break" in the structure. These labels are NEW, NEW-REB, CONC*, REASONp, and CONCp; thus the "breaking" segments of the text are 3, 5, 7, 10, 14, 16, and 18. In the database scenario described above, we could now for these segments look at annotations of connectives, thematic structure, and sentential information structure in order to check whether a "break" in the argument is signalled in some way or another at the linguistic surface. (But recall that we are still in the step of the pilot study and thus do not have enough argument structure annotations at hand yet.)

For our sample text, when inspecting the topic development of the text, we indeed find motivations for each "breaking" segment listed above. 3 introduces the general topic *Olympic games*, while 5 shifts it to *Berlin* (subject and topic). This topic is maintained in 6 with the phrase *the capital*, while 7 shifts subject and topic to a different city, *Barcelona*. In 10, we find a preposed conditional clause that can be analyzed as a frame-setting topic;[4] furthermore the clause comes back to *Berlin*, closing off the excursion to Barcelona. While 13 discusses the 2000 venues, 14 via topicalisation of the prepositional phrase moves to Berlin's new application. Finally, 16 signals a shift with the connective *and even if*, which introduces a clause that, like 10, can be seen as frame-setting for the subsequent clause.

[4] 'Frame-setting' topic on the level of sentential information structure is often used to characterize locative or temporal phrases introducing the topic of the sentence; here, we are extending the term to a similar class of situations on the discourse level.

A complementary step of analysis is not governed by the segments' order in the text but by traversing instead the support graph and checking for adjacent nodes that correspond to alleged long-distance relationships in the text. Consider, for example, the link between nodes 6 and 18 in the graph. Though the label sequence already identifies 18 as a "break" in this case (because it is reached from the counter-rebuttal in 17), we are in general not guaranteed to locate every gap of this kind with the sequence labels. In this particular example, the 6-18 link can be explained as follows: 6 quite explicitly states the position of the author relatively early in the text; 18/19 is a metaphorical ending of the text that more or less directly, and here quite informally, sums up that position. (We encountered figures like this quite often in our corpus.) Notice that the reader is invited to make this connection not only on the grounds of "deep" meaning but also by lexical cohesion: 6 introduces the *running* metaphor *(baton)* that 18 eventually returns to by evoking the *starting block* image.

4 Conclusion

The multi-level representation of discourse information allows for systematically uncovering relationships such as the one between argument structure and aspects of surface realization, as outlined in Section 3. In principle, the method can be employed in two different ways.

1 - We can explicitly pose specific queries to a database in order to test a particular research hypothesis we are already entertaining. An example is the study of (Chiarcos & Krasavina 2005), who used our commentary corpus to set the levels of referential and rhetorical structure into correspondence and check whether the notion of "rhetorical distance" (as proposed for example in the "Veins Theory" of Cristea et al. 1998) influences the author's decisions on pronominalization. As a technical infrastructure for this type of work, we have developed a linguistic database *(ANNIS)* and a data exchange format together with conversion scripts for mapping from a range of widely-used annotation tools to the exchange format and to ANNIS; see (Chiarcos et al. 2008).

2 - The other approach is to perform traditional data mining techniques on a corpus with multiple annotation levels in order to discover patterns that we did not explicitly search for or anticipate – a step we can call "annotation mining". To support this, we provide a mapping from our database to the input format of the WEKA toolkit (Witten & Frank 2005), which offers a range of modules implementing both supervised and unsupervised machine learning techniques.

As such, multi-level representation and analysis for text corpora is designed as an alternative to accounts that aim at encoding "the" discourse structure in a single framework, e.g., RST or SDRT. In (Stede to appear), it is demonstrated that the multi-level approach can eliminate a variety of ambiguities that are inherently encoded in RST-style analyses. We believe that text corpora should be maximally useful for a wide range of purposes, and the idea of first distinguishing the different realms of information from one another, and then flexibly combining them for a given research question or application can be very helpful. At the same time, the technique gives rise to new methodological questions, which we mention here as issues for future work: How should possible dependencies between annotated levels be dealt with? For instance, given a level of sentence syntax, the NPs can conveniently be used as "markables" for a level of coreference annotation. For annotation mining, however, dependencies of this kind need to be taken into consideration when deriving conclusions from the corpus. Similarly, the approach allows for representing competing analyses on the same level, e.g., when two annotators produce alternative accounts of argument structure. When one of these levels is combined with some other level for deriving new information, the presence of the alternative account should not be neglected.

References

Nicolas Asher & Alex Lascarides. *Logics of Coversation.* Cambridge University Press, Cambridge, 2003.

Margareta Brandt & Inger Rosengren. Zur Illokutionsstruktur von Texten. *Zeitschrift für Literaturwissenschaft und Linguistik*, 86:9-51, 1992.

Christian Chiarcos & Olga Krasavina. Rhetorical distance revisited: a parameterized approach..In *Proceedings of the Workshop on Constraints in Discourse (CiD),* Dortmund, 2005.

Christian Chiarcos, Stefanie Dipper, Michael Götze, Julia Ritz & Manfred Stede: A flexible framework for integrating annotations from different tools and tagsets. In: *Proceedings of the First International Conference on Global Interoperability for Language Resources (ICGL-08)*, Hongkong, 2008.

Dan Cristea, Nancy Ide & Laurent Romary: Veins Theory: A Model of Global Discourse Cohesion and Coherence. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the ACL (COLING/ACL- 98)*, Montreal, 1998.

James B. Freeman. *Dialectics and the Macrostructure of Argument.* Foris, Berlin, 1991.

Katja Jasinskaja, Jörg Mayer, Jutta Boethke, Annika Neuman, Andreas Peldszus, & Kepa Joseba Rodríguez. *Discourse tagging guidelines for german radio news and newspaper commentaries*. Ms., Universität Potsdam, 2007.

William Mann & Sandra Thompson. Rhetorical Structure Theory: Towards a functional theory of text organization. *TEXT*, 8:243-281, 1988.

Chris Reed. Representing dialogic argumentation. *Knowledge-based systems,* 19(1):22-31, 2006.

Manfred Stede. The Potsdam Commentary Corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*, pages 96-102, Barcelona, 2004.

Manfred Stede. *Korpusgestützte Textanalyse.* Narr, Tübingen, 2007.

Manfred Stede. RST revisited: disentangling nuclearity. In Catherine Fabricius-Hansen & Wiebke Ramm, editors, *'Subordination' versus 'coordination' in sentence and text – from a cross-linguistic perspective*. John Benjamins, Amsterdam, 2008.

Manfred Stede: Disambiguating rhetorical structure. To appear in *Research on Language and Computation.*

Stephen Toulmin. *The Uses of Argument.* Cambridge University Press, Cambridge, 1958.

Egon Werlich. *Typologie der Texte*. Quelle und Meyer, Heidelberg, 1975.

I. Witten & E. Frank. *Data Mining: Practical Machine Learning Tools*. Morgan Kaufmann, San Francisco, 2005.