DHQ: Digital Humanities Quarterly
2022
Volume 16 Number 3

Algorithmic Close Reading: Using Semantic Triplets to Index and Analyze Agency in Holocaust Testimonies

Lizhou Fan <lizhouf_at_umich_dot_edu>, University of Michigan
Todd Presner <presner_at_ucla_dot_edu>, UCLA

Abstract

The following article presents a digital humanities exploration of indexing and analyzing expressions of agency in Holocaust testimonies. Using a set of text analysis methods to identify, classify, and visualize "semantic triplets," we show how attention to agency complements and extends conventional approaches to indexing. Our examples come from two corpora of Holocaust oral histories: the first were conducted in Displaced Persons camps in 1946 by an interviewer named David Boder; the second were conducted by the USC Shoah Foundation Visual History Archive in the 1990s. We focus on two salient testimonies from each corpus in order to describe the methodology and what the analysis of agency can contribute to the writing of "microhistories" of the Holocaust. Building on semantic web analyses, the methods provide a groundwork for the development of a graph database to search testimonies by agency and thereby provide historical insights about what people report they did and what was done to them.

1 Introduction

Oral histories conducted with survivors are infinitely rich sources for understanding the complex events of the Holocaust. As first-person narratives, oral history interviews with Holocaust survivors provide vital, eyewitness testimony about the rise of Nazi Germany, the violent acts of genocide, and reflections on its aftermath. The "era of the witness," as Annette Wieviorka calls it, coincides with the emergence of institutional recording initiatives, such as those at Yad Vashem, Yale's Fortunoff Video Archive, the US Holocaust Memorial Museum (USHMM), and the USC Shoah Foundation Visual History Archive [Wieviorka 2006]. Altogether, these archives have recorded well over one-hundred thousand hours of testimony and preserved hundreds of thousands of individual stories. 1

Because of the sheer size of these collections, indexing systems are critically important for accessing their content. As such, all of these collections can be searched effectively by names, dates, events, places, organizations, and various subject headings. This is because "noun-based ontologies" form the taxonomic structure of most indexing systems and finding aids. Attuned to the objective realm of what was said, these ontologies help make the general content of an oral history discoverable and accessible to researchers and the broader community of listeners. When users enter keywords into a textbox, search and retrieval works because the content is consistently and thoroughly described by a standardized, controlled vocabulary [Presner 2016]. 2

Not surprisingly, this is also the approach of most entity-recognition software packages, such as OpenNLP [Apache Software Foundation 2017] and SpaCy [Explosion AI 2020], which are very successful at identifying things said, such as proper names, organizations, geographic locations, historical events, dates, units of time, quantities, and other numerical units, such as percentages and times of day. But one of the fundamental limits is the focus on extracting isolated elements (primarily nouns) rather than larger narrative units that describe what people did and what was done to them. The purpose of this paper is to ask: In what ways can oral histories be mined computationally to index how survivors narrate behaviors, actions, reactions, assessments, and modes of self-orientation and personal agency? This is vitally important in Holocaust testimonies since agency is often described within networks of violence and coercion. 3

This paper explores a method for extracting, characterizing, and interlinking “semantic triplets” to index and analyze vectors of action and agency, rather than just discrete topics, themes, or isolated nouns. In so doing, we complement conventional indexing to gain a richer, nuanced understanding of the events, agents, and actions described in a testimony.

A semantic triplet is a grammatical unit consisting of *Subjects*, *Relations*, and *Objects*. *Subject*s are units consisting of nouns, primarily people, pronouns, and proper names, as well as direct modifiers of the subject. *Relations* are verb units including one or more verbs, related prepositions, and modifiers like adverbs. *Objects*, both direct and indirect, have a broader range of parts of speech and include regular noun units, adjectives, adverbs, and more. For example, “I removed the yellow patch”[1] is a semantic triplet expressed as active speech (something the subject did), while “I was shipped to Auschwitz” is a semantic triplet expressed as passive speech (something done by an unspecified agent to the subject). By linking semantic triplets with descriptive, keyword metadata, we hope to show how new patterns of agency and previously unrecognized “microhistories” [Zalc and Bruttman 2016] can be highlighted and discovered within oral histories. Rather than considering our method as a reduction or simplification of the text into discrete units of data, we argue that semantic triplets form a paratext that highlights the richness of the testimonial narrative by adding to existing metadata and methods of indexing. 4

2 Indexing Approaches in Two Corpora of Holocaust Testimonies

2.1 Corpus #1: David Boder’s 1946 Interviews with Displaced Survivors

The psychologist and linguist David Boder has a unique place in the history of Holocaust testimony collection because he was the first person to record the voices of displaced survivors in their own words after the end of World War II.[2] In the summer and early fall of 1946, he conducted over a hundred interviews with Jewish survivors and other displaced people in nine different languages (mostly in German and Yiddish, but also in English, Russian, Polish, French, Lithuanian, Latvian, and Spanish), using a relatively new technology called a wire recorder.[3] He carried out his interviews in Displaced Persons (DP) camps and other safe houses in France, Switzerland, Italy, and Germany, mostly with Jewish survivors of concentration camps and slave labor camps, as well as with a smaller group of non-Jewish refugees consisting of Mennonites and other Christians who had fled Soviet territories. The interviews range in length from about 15 minutes to over four hours, with most about one-hour in duration. He asked over 16,000 questions to his subjects in an attempt to understand the historical events of the Holocaust and learn about his interviewee’s personal story of loss and survival.[4] Today, Boder’s interviews are available online, as audio files in their original languages as well as in English translation, through the Illinois Institute of Technology’s “Voices of the Holocaust” project.[5] 5

After the interviews were recorded, Boder spent nearly ten years preparing a 3,194-page anthology of the survivor reports, which he called *Topical Autobiographies of Displaced People*.[6] It consisted of typewritten translations, which he marked-up and indexed, of 70 of the testimonies. He developed the first indices for characterizing the experiences of Holocaust survivors, including a “subjects and situations” index with more than 300 entries, a “geographic and ethnic” index with over 400 entries, a “persons and organizations” index with close to 300 entries, and a “trauma inventory” enumerating 46 types of trauma and violence experienced by survivors. His “subjects and situations” index represented the first controlled vocabulary for indexing Holocaust testimonies and included shared experiences like deportations, selections, mistreatment, and killings as well as personal experiences such as food deprivation, thirst, and screaming. 6

Our analysis below will focus on two of these testimonies: The first is an interview he conducted on September 26, 1946, with a 34-year-old, Polish-Jewish woman named Anna Kovitzka (given the pseudonym “Anna Kaletska” by Boder). Kovitzka lost most of her family, including her young baby, who she had given to a Christian Polish woman in the hopes that the child would be able to survive. That Polish woman was later denounced by a neighbor, and the child was killed. The hour-long interview, spoken in Yiddish, covers about eight years of her life and moves between her birthplace of Kielce, escaping the Grodno ghetto, being deported to Auschwitz, and, finally, at the end of the war, being driven back into Germany, and ending up after the war in Wiesbaden, where she is interviewed. The second interview is with a German-Jewish survivor named Jürgen Bassfreund, who was also deported to Auschwitz, where he worked in a 7

forced labor factory. Conducted in Munich on September 20, 1946, Bassfreund describes the rise of Hitler in Germany, the deportation and death of his mother, and the horrific conditions he endured while being transferred between multiple concentration camps. We chose this particular interview because Bassfreund – who later changed his name to Jack Bass – was re-interviewed by the USC Shoah Foundation in 1997, in his home in Alabama.

2.2 Corpus #2: The USC Shoah Foundation Visual History Archive

Containing about 55,000 audiovisual testimonies, the Shoah Foundation's Visual History Archive (VHA) is the largest digital archive of Holocaust and genocide testimonies in the world. The videos range in length from several minutes to more than ten hours, although most are between one and three hours in duration. Indexers trained by the Foundation tagged the videos with keywords by linking one-minute segments of video to one or more search terms in its 60,000+ word, hierarchical thesaurus. In addition, transcripts are currently being created for many of the testimonies in their original languages as well as in English translation.[7] 8

Using a tool called a "Video Indexing Application," the majority of the testimonies were automatically separated into discrete, one-minute segments, and an indexer was prompted to assign a keyword from the thesaurus to the segment (although not every segment had to have an indexing keyword). The VHA derived its indexing protocols from the National Information Standards Organization's Z39.19 standard for the construction, format, and management of monolingual controlled vocabularies. While useful for identifying topics, people, places, and events for general searching, the index tends to obscure individual actions and agency, not to mention aspects of performative delivery, subjective experiences, and expressions of questioning, doubt, and possibility [Presner 2016]. Even though the aim of the index is objectivity, it is important to underscore that a human listener decided what to index and what not to index; a human listener decided what indexing term to use and what indexing term not to use; and a human listener decided if a given narrative segment should be described by a keyword or not. What is missing in the indexing system's "philosophy of pursued objectivity" [USC Shoah Foundation 2006] are the specific semantic expressions used by survivors to describe actions, give evaluative assessments, provide orientation, and tell about other types of agency and activities, including subjective experiences and everyday behaviors, reactions, and acts of resistance. 9

In addition to the Jürgen Bassfreund/Jack Bass re-interview, our analysis will focus on the testimony of Erika Jacoby. Jacoby, a Hungarian Jew deported to Auschwitz with her family in the Summer of 1944, survived the camp with her mother. Her family was among the hundreds of thousands of Hungarian Jews deported to Auschwitz at the height of the mass murder operations at Auschwitz. She was separated from her grandparents and other members of her family at the ramp of Auschwitz-Birkenau and never saw them again. She was able to stay together with her mother, and both were selected for forced labor at Auschwitz. Her testimony was recorded in her home in Los Angeles in 1994. Like the testimony of Bass, the interview follows the structure of a chronological life-story, beginning with childhood, experiences during the War and Holocaust, liberation, and eventual immigration to the United States. 10

The four testimonies discussed in this article are intended to provide a proof-of-concept for a method to identify and index expressions of agency. Our hope is that the methods can be used to complement conventional, subject-based indexing and open new possibilities for cross-corpora analysis. 11

3 Methodology

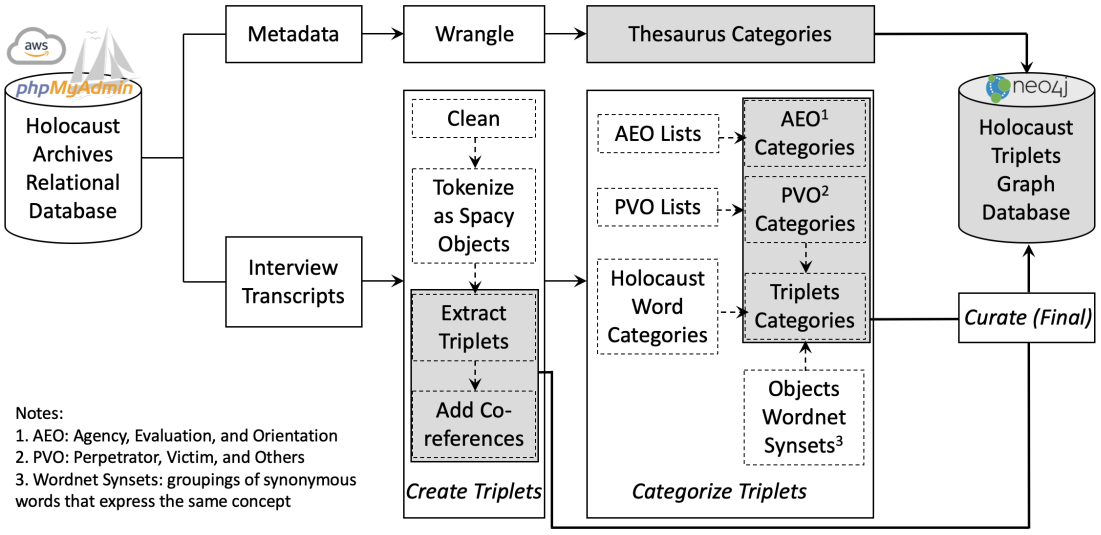


Figure 1. Processing and Categorizing Triplets

3.1 Semantic Triplets Extraction

At the highest level, the aim of our research method is the transformation of a Holocaust archives relational database (keywords linked to testimony segments) into a Holocaust triplets graph database (networks of semantic triplets derived from the testimonial transcripts). As shown in Figure 1, we use both interview transcripts and their metadata, while our main contribution, as discussed in this article, is the identification, extraction, and categorization of semantic triplets from the transcripts. Our process can be summarized as follows: First, we identify and extract triplets algorithmically from the cleaned transcripts; then, we use text mining methods to assign categories and cluster the triplets; finally, the results are manually reviewed (human check, correction, and curation). After that, they can be stored, along with the original metadata, in a Tableau dashboard or eventually a graph database for querying, indexing, and visualization. We will introduce these core methods below. 12

Triplestore and RDF store are databases optimized for storing and retrieving semantic triplets ([Rusher 2006]; [Peter 2001]). In general, triplets take the form of subject-predicate-object expressions and can be searched through network relations between entities. Digital humanities scholars have used semantic web methods, such as linked data, to facilitate interoperable research structures to analyze social or cultural datasets [Hyvönen 2020] as well as to model scholarly processes in the humanities by exploring the productive tension between formally structured, semantic data and humanities scholarship [Bradley and Pasin 2017]. Inspired by both the database design of Triplestore and RDF store and their humanistic applications, we have developed our own process of extracting semantic triplets from spoken testimonies. Our approach is best situated in the field of information extraction for domain specific documents (in our case, Holocaust and genocide testimonies), while the methods and ideas are replicable in other domains such as oral histories more broadly, autobiographies, and personal narratives. 13

Traditionally, researchers proceed by creating open information extractors using token related patterns ([Yates et al. 2007]; [Fader et al. 2011]) or dependency parsing related features ([Wu and Weld 2010]; [Mausam et al. 2012]). With the advancement of machine learning technologies, we now see a trend of switching the research focus from extracting informative and broad coverage semantic triplets to producing canonical triplets which are linguistically coherent [Angeli et al. 2015]. This trend of research tends to focus on improving the extractors' accuracy, and, to date, there are few applications that use semantic triplets to analyze spoken, personal narratives, which resist formalized structures. Our semantic triplets extraction method is an automated, rule-based process to aid close reading and focuses on extracting expressions of agency and statements that speak to memories of individual and collective actions. 14

This triplet extraction method leverages a combination of chunk parsing, also known as partial parsing, and domain specific dictionaries. It focuses on the lower-level chunks of nouns and/or adjectives (NAP) and verbs, which are groups 15

of connected but non-recursive tokens with different levels in syntactic structures of sentences.[8] We convert the basic linguistic representation of the chunks' rules with regular expression from the NP and VP expressions of Abney (1996)'s rules as:

$$NAP \rightarrow D?A*N*N|D?A*A$$

and

$$VP \rightarrow V_{-tns}|AuxV_{-ing}$$

where D means determiners, A means adjectives, N means nouns, V_{-tns} means the variation of a verb depending on its tense, Aux means auxiliary, and V_{-ing} means the present participle of a verb. Using the natural language processing tools in SpaCy, we further extend these rules for chunks using the built-in Noun Chunk function and a customized Matcher of verb chunks.[9] Our extension includes more parts of speech as the possible components around core nouns, verbs, and adjectives, and we link the chunks together to check if we fully process the sentences.[10]

For example, Figure 2 shows a parsing tree with the different parts of a triplet. In the sentence "I was clinging to the window," we detect the active agency of window clinging with "I" as the subject noun chunk and "the window" as the object noun chunk. In this case, the verb chunk includes the verb, an auxiliary verb, and a preposition. Using this extraction method, in our four example testimonies, we are able to extract more than 3,000 triplets. Some sentences may contain multiple triplets, such as this sentence: "I went with her down to the door, and there I stood across the street, hidden in the gate, and I saw how my child was lying on the snow." In this example, four triplets were extracted: "I went with her," "I stood across the street," "I hidden in the gate," and "my child was lying on the snow." 16

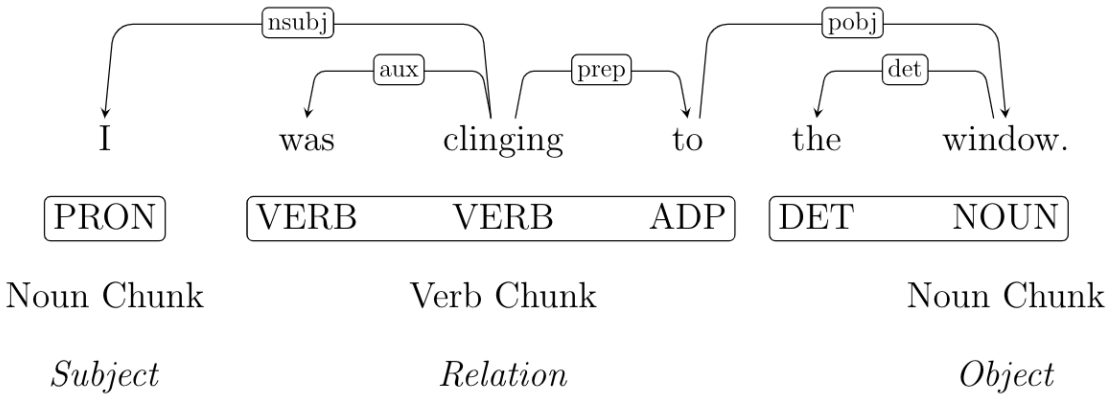


Figure 2. A Parsing Tree with Lexical Chunks that Form a Triplet

When parsing longer sentences, we always include, in a separate column for each triplet, what we call "context terms." We define the context terms to be all the related lexical chunks other than those included in the extracted triplet. In the first triplet above ("I went with her"), the contextual chunks are the following words: the door, the street, the gate, my child, the snow. The context is especially important when we want to understand an action embedded in a complex sentence or create network diagrams from the triplets.[11] 17

3.2 Characterization: Action, Orientation, and Evaluation (AEO) Categories with Subcategories

One of our methods for characterizing semantic triplets is derived from Labov and Waletzky's [Labov and Waletsky 1997] social-linguistic model of narratives, where clauses are characterized by temporal organization (the order in which the subject narrates events and actions in the story), evaluative description (personal assessments made by the 18

narrator), and contextual orientation (usually information provided by the narrator that helps orient the listener). Statements of evaluation provide judgment about a person or situation, while statements of orientation provide contextual or relational information. Various scholars have developed tools or solved related problems using the Labov model: For example, Swanson et al. [Swanson et al. 2014] created a part of speech (POS) based algorithm for clause type detection and labelling based on this model; and Saldias and Roy [Saldias and Roy 2020] used this model to train a classifier to analyze and compare spoken personal narratives. Although these applications to characterizing narratives focus on the clause level, the Labov model is also applicable to our research because triplets are derived from clauses and need to retain semantic information.

Not unlike the use of text mining to study literary "styles" [Jockers and Underwood 2015], we use our text mining methods to classify triplets by similar types of speech in order to discover relationships between different types of expressions of agency in oral histories. We divide descriptions of agency into four sub-categories: active, passive, coercive, and speculative speech. Active and passive speech are, respectively, actions done by the subject of the triplet and actions done by others to the subject; coercive speech may have an active structure but a passive meaning since it represents something the subject was forced to do (sometimes without mentioning the person or group responsible for the coercion); and, lastly, speculative speech, often expressed through modal verbs, represents an uncertain or imaginary action, or a statement of possibility, desire, ability, or futurity. While these categories are not perfect or definitive, they help to identify the general kinds of speech describing actions. We summarize and provide examples for the two-level AEO characterization system in Table 1.

Categories		Examples
Agency	Active	I removed the yellow patch.
	Passive	I was shipped to Auschwitz.
	Coercive	My father had to leave his home.
	Speculative	They could consummate the terrible deed.
Evaluation		I am not ashamed.
Orientation		I remembered a Christian woman.

Table 1. Two-level AEO Characterization System with Examples

The Triplets AEO Algorithm implements the above process as a basic characterization system. In SpaCy, the algorithm takes tokens consisting of the Relation Verb Chunks, R, and Object Noun Chunks, O. For each token in both R and O, there is retrievable syntactic information including (but not limited to) a part of speech (POS) tag, a dependency tag, and the lemma form of the word. We then put together this lexicon- and rule-based mechanism as shown in Appendix I, Algorithm 1.

3.3 Characterization: Perpetrators, Victims, and Others (PVO) Categories

The tripartite approach to characterizing victims, perpetrators, and bystanders has been around for decades in Holocaust historiography [Hilberg 1992]. While not absolute or unproblematic, we strive to disambiguate subjects and objects in triplets when they are identified as people or groups. To do so, we first used the Entity Recognition pre-trained models in SpaCy to determine the category of a NOUN (people, organizations, geographic locations, and so forth). We, then, manually classified the detectable entities, where possible, into victim and perpetrator groups. For example, the phrases "the German Army," "the SS," or "the Gestapo" are detected as organizations and marked as Perpetrator, whereas the interviewee ("I" and "we") as well as familial units (parents, grandparents, sister, brother, and so forth) are marked as Victim. Pronouns, however, were not disambiguated. This classification method will always be incomplete, as many subjects and objects do not unequivocally belong to the victim or perpetrator group. The "gray zone" contains bystanders, aid-providers, "implicated subjects," and people who moved across categories and do not easily fit the binary ([Rothberg 2019]; [Luft 2015]). These and other such subjects and objects are also extracted but not categorized, as they demand human scrutiny of context and, indeed, prompt further questions and future research about the range of

agents in an oral history. For the ones that are certain, we created lists of victim and perpetrator lemmas, which provided the base forms of the entity words in the triplets.

4 Results and Applications

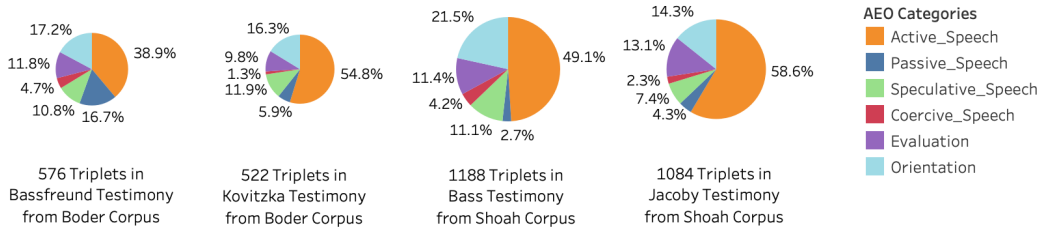


Figure 3. Distribution of Semantic Triplet Categories in Four Testimonies

In Figure 3, we show statistical summaries of the triplets in each of the four testimonies. The triplets were algorithmically generated from the transcripts, assigned categories by speech type, and evaluated for accuracy. The size of each pie chart is scaled to the number of triplets extracted for each testimony. Boder's interviews are shorter than the Shoah Foundation interviews (and, of course, end in 1946, rather than the 1990s). Due to the large number of triplets, we needed to develop ways to categorize and cluster the results. This entailed categorizing the triplets into six types of speech, assigning Victim and Perpetrator identities to certain subjects and objects, and experimenting with using a Wordnet-based characterization of the objects to cluster them by lemma and categories.[12] For our initial version of the triplet extraction process, 6.5% of the triplets were manually corrected due to incorrect or incomplete subjects, verb relations, or objects. In addition, about 14% were manually removed from the original output because the triplet did not make sense grammatically or was misidentified as a triplet. The challenges and shortcomings of the triplet extraction process are discussed in section 5.2. To better analyze the triplets and their characterization, we created an interactive Triplets Dashboard, shown in Figure 4, which brings together the corrected triplets output, an object lemmas bubble plot, and multiple other filters. 23

As a finding aid, the dashboard helps us query and sort the triplets by interview, subject, object, speech type, and more. In the dashboard, a search for semantic triplets can be initiated on any term, category, or type of speech both within a given testimony or across a group of testimonies. In the screenshot, we see the triplets around Bassfreund's description of his deportation from Berlin to his selection for forced labor in Auschwitz. They evidence a wide-range of speech acts, including active speech about other people's agency ("they gave us six slices" and "they placed a can"), passive speech ("we were put on trains" and "we were driven out of the cars"), coercive speech ("we had to sign affidavits" and "we have to leave Germany"), speculative speech ("I shall never forget those screams"), evaluation ("it was uncomfortable" and "the screams were terrible"), and orientation ("I have forgotten something" and "the people had a premonition"). Looking further at the contextual information for the coerced action of being forced to sign affidavits, for example, we see the following terms: "our disloyalty," "Germany," "account," and "hostility." The original sentence can be seen on the far right of the triplet output, as translated by Boder: "Before being sent away, we had to sign affidavits that we have to leave Germany on account of our disloyalty and hostility to Germany."[13] 24

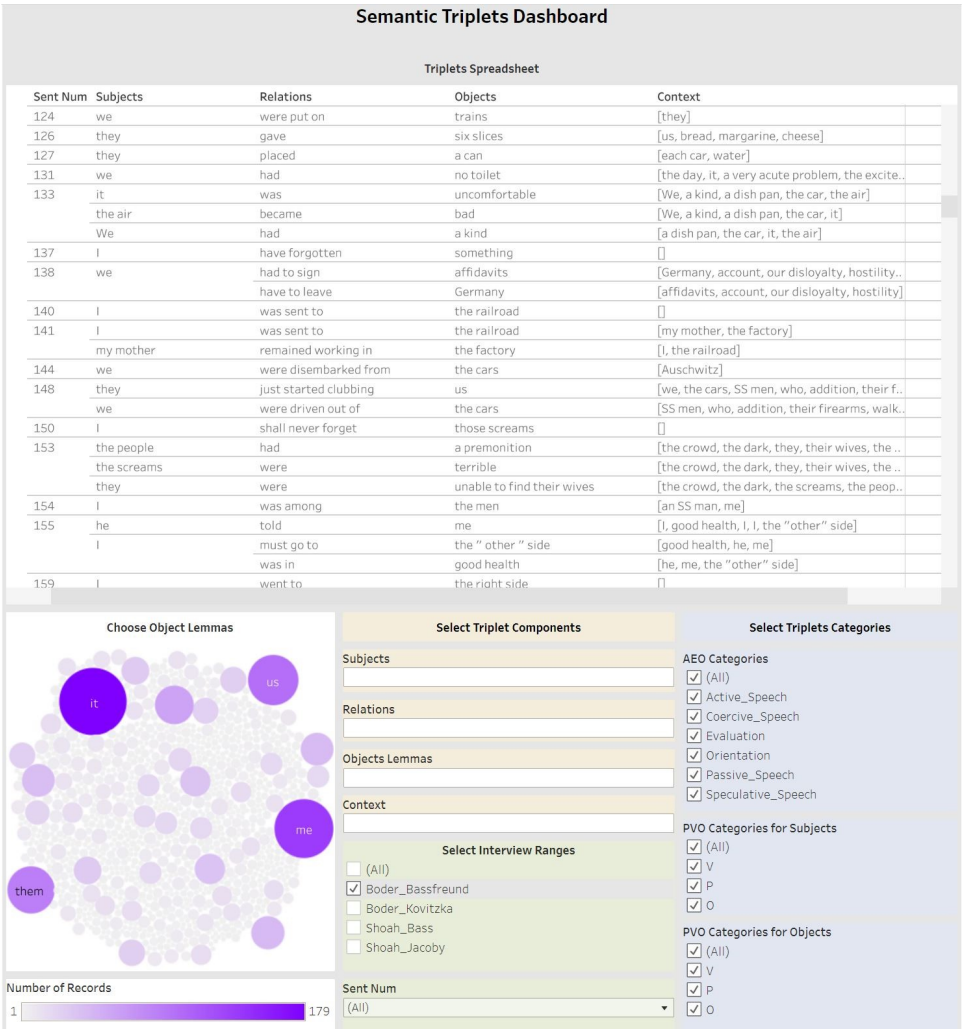


Figure 4. Semantic Triplets Dashboard in Tableau (with simplified objects and context terms). Selection from Boder's interview with Jürgen Bassfreund

Another way triplets can be used is to identify instances of different types of speech within a testimony. For example, leading up to and following his description of being deported to Auschwitz, we can visualize all the passive and coercive statements made by Bassfreund in which "I" or "we" is the subject. As shown in Figure 5, there are over 50 such instances of passive and coercive speech described in this section of his interview with Boder (about ten type-written pages). Bassfreund emphasizes, with precise language, what was done to him and the Jewish community by the Nazis (we were surrounded, taken, moved, driven out, loaded on, led to, rubbed with, shoved into, sent into, transferred to) as well as what was specifically done to him (I was forcibly taken, assigned, sent, transferred).[14]

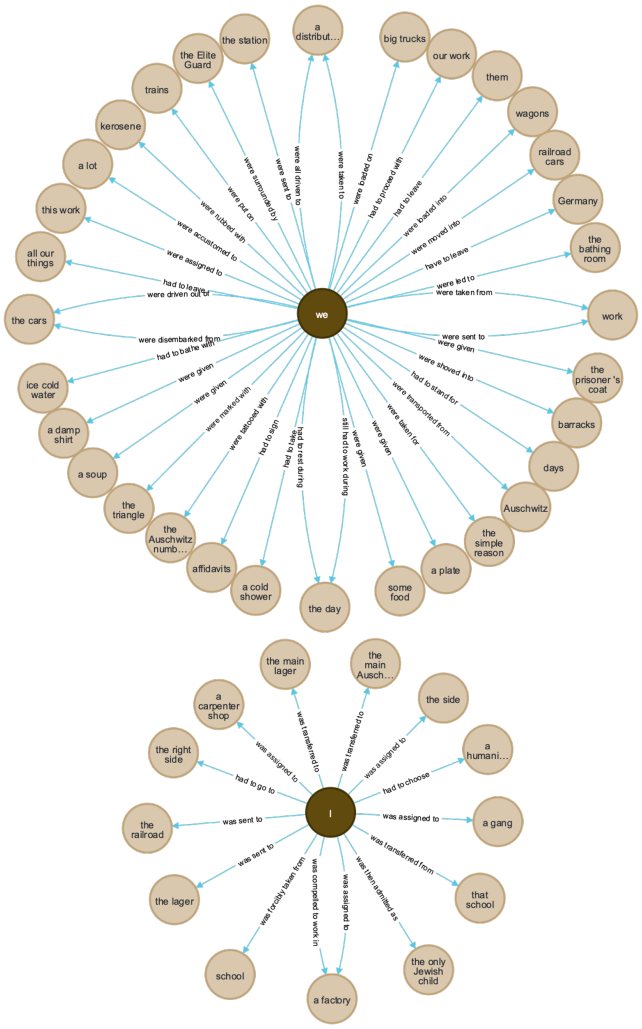
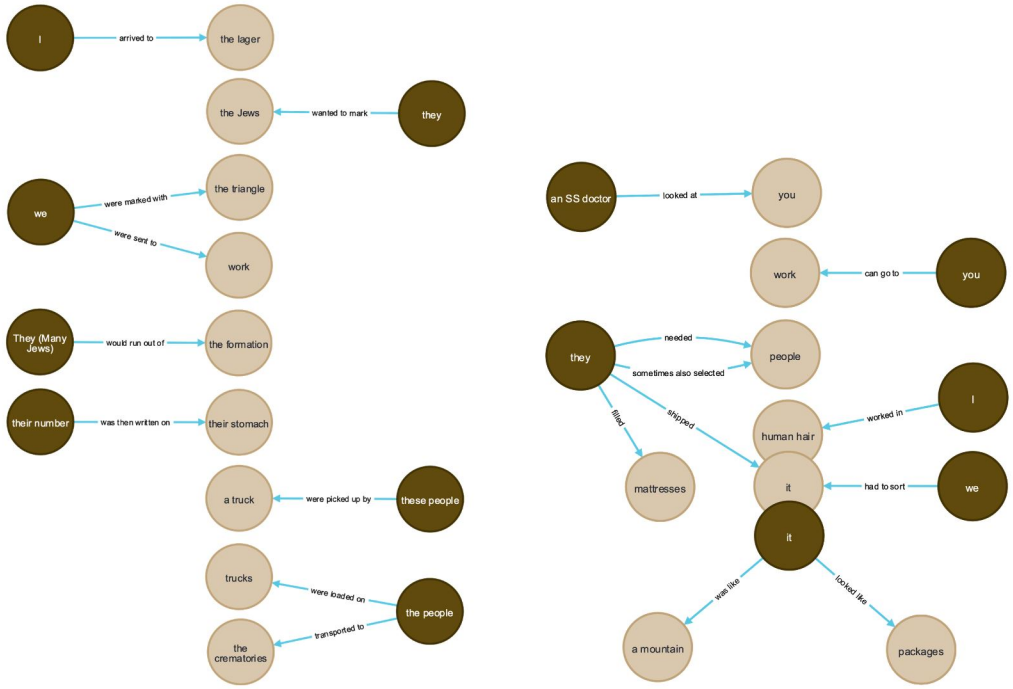


Figure 5. Selection of Bassfreund's Passive and Coercive Speech (Boder interview), with "we" and "I" as subject, respectively

We can now show our approach to semantic triplets extends and deepens topical indexing by pairing the triplets with the human-created indexing terms. As developed by both Boder and the Shoah Foundation, we see a significant continuity in terms of topics and themes (places, time periods, people, conditions, and events) used to index the testimonies. For instance, on pages 292-293 of Bassfreund's testimony in *Topical Autobiographies*, Boder indexed the testimony with the following terms: bathing, clothing, hair cutting and shaving, professional criminals, confiscations, punishments, and tattoos. Similarly, in segment 25 of Bass' testimony, we see the following indexing terms from the Shoah Foundation: Auschwitz I (Poland: Concentration Camp), camp forced labor, camp latrines, camp living conditions, camp selections, Poland 1941 (June 21) - 1945 (May 7), prisoner hair cutting forced labor, transfer from Auschwitz III-Monowitz (Poland: Concentration Camp), and transfer to Auschwitz I (Poland: Concentration Camp). While these topics certainly give us guidance for finding the overarching places, dates, conditions in captivity, movement, and forced labor, the indexing is not at the level of sentences or clauses and, therefore, we do not really know what Bassfreund/Bass did or what was done to him. As we see in Figure 6, we produce a corresponding network of just a few of the triplets derived from these specific parts of his testimonies. The interrelated set of triplets provide information about un-indexed subjects and agents (both victims and perpetrators), un-indexed actions (written on, picked up by, worked in, filled, sent), and un-indexed objects (the formation, their stomach, a truck, the crematories, human hair, mattresses, packages, and a mountain). Altogether, they point us toward ways of using the extracted semantic triplets to read, mark-up, and index testimonies by including microhistorical accounts of actions by both victims and perpetrators.



(a) Bassfreund (interviewed by Boder) (b) Bass (interviewed by USC Shoah Foundation)

Figure 6. Selected Examples of Triplets related to Indexing Categories in Two Testimonies. (a) Bassfreund (interviewed by Boder) (b) Bass (interviewed by the Shoah Foundation)

In the next section, we discuss some of the applications of the triplets to identifying acts of resistance, detecting unindexed stories, and analyzing larger semantic networks. 27

4.1 Microhistorical Acts of Resistance

While large-scale acts of resistance (such as the Warsaw Ghetto uprising or the Sonderkommando rebellion in Auschwitz) are rightly known, there has been comparatively less appreciation of the range of everyday, seemingly ordinary acts of resistance by individuals. One historian who is working at the vanguard of this reappraisal of Jewish agency is Wolf Gruner, who has delved into numerous police and municipal archives in Germany to uncover the wide-range of acts of defiance, opposition, and protest by Jews living in Nazi Germany ([Gruner 2011]; [Gruner 2016b]). In his studies of “microhistories,” he has unearthed thousands of examples of Jews mounting both formal and informal protest, ranging from the filing of petitions and government complaints to outright acts of defiance, such as sitting on benches marked “Aryan only” or refusing to use the forced middle name of “Israel” or “Sara,” to physical and armed resistance against Nazi officials [Gruner 2016a]. For our purposes here, we are interested in using our computational method of identifying networks of semantic triplets to find underappreciated or largely unknown acts of everyday resistance within Holocaust testimonial narratives. Our argument is that triplets point to a wide-range of small-scale actions that would otherwise go unremarked, since they do not, generally speaking, rise to the level of what might be indexed as “acts of resistance.” Oftentimes, these actions are not even considered to be “indexable content,” and thus they are extremely difficult to locate. 28

4.1.1 Example 1: Anna Kovitzka

Interviewed by Boder in 1946, Anna Kovitzka’s story is particularly poignant, as she describes the violent separation and loss of her family and, in particular, her efforts to save her baby girl who was born in the Grodno ghetto. With her husband, she managed to escape the ghetto in the hopes of finding someone to care for her child. Shortly after giving up the child to a Christian woman who promised to protect her, Kovitzka was deported to Auschwitz. The child, 29

unfortunately, did not survive as the Christian woman was denounced for trying to protect a Jewish baby. Examining the indexing terms used by Boder to mark-up her testimony, we see a wide-range of themes and topics, including: geographic locations (Wiesbaden, Kielce, Grodno, Slonim, Lvov, and Auschwitz), living conditions and objects (epidemics, starvation, thirst, illness, sleeping accommodations, and clothing), people (children, Gestapo, Prisoners-of-War, family, and Gypsies), and events (looting, childbirth, burials, flight, killings, escapes, work, bathing, rebellion, and appells or roll calls in camps). The last group of events certainly raises questions about actions, but without further analysis, we cannot know who was looting, escaping, working, rebelling, or even giving birth to a child. Moreover, without delving into the testimony, we cannot uncover specific vectors of agency (such as who did what, how did they do it, and to whom).

30 By examining the semantic triplets around the topical indexing term *escapes,* we can illuminate the specific agency described by Kovitzka in this part of her testimony (Table 2). The triplets indicate the actions of Kovitzka and her husband to escape the ghetto in order to save their child: “went over” the wires, “set up a chair,” “raised” the wire, “went out” on the street, “removed” the yellow patch, and, finally, “went down” the street. In each case, we see active speech describing a set of actions that occurred in the context of escaping the ghetto, even after she reports that sixteen Jews were killed at the gate in a first attempt to escape. While this section of the text is indexed by Boder as *escapes, curfew, fences,* and *yellow star,* the actual acts of escaping the ghetto through the wires, removing the yellow Star of David, and going down the street represent acts of resistance and agency that can be successfully detected and added to the index through computational text analysis attuned to semantic triplets.

Subjects	Relations	Objects	Context and Coreference
Sixteen Jews	fell at	the gate	first attempt
I	went over	the wires	second attempt, my man, me, the child, her
my man	handed	the child	I, the wires, the second attempt, me, I, her
My man	set up	a chair	me
He	raised	one wire	Coreference: He = my man
He	handed	the child	me, Coreference: He = my man
I	went out in	the street	
I	removed	the yellow patch	
I	went down	the street	

Table 2. Anna Kovitzka Triplets (around *escapes*)

31 Using the dashboard to search on the object *child* in Kovitzka’s testimony reveals a number of triplets that show her attempts to save her child and, ultimately, the child’s fate: “I must save my child,” “My man handed the child,” “she has taken my child,” “she had picked up the child,” “I don’t have my child,” “she kept the child,” “she loved the child,” “my child was lying on the snow,” and “they buried the child.” The tragedy of Kovitzka’s loss of her child forms the traumatic core of her testimony and, yet, there is barely any indication of this from the indexing terms. The creation of a semantic network around the term *child* would help us locate and appreciate these critical actions and thus begin to index the testimony attuned to expressions of agency (Figure 7).

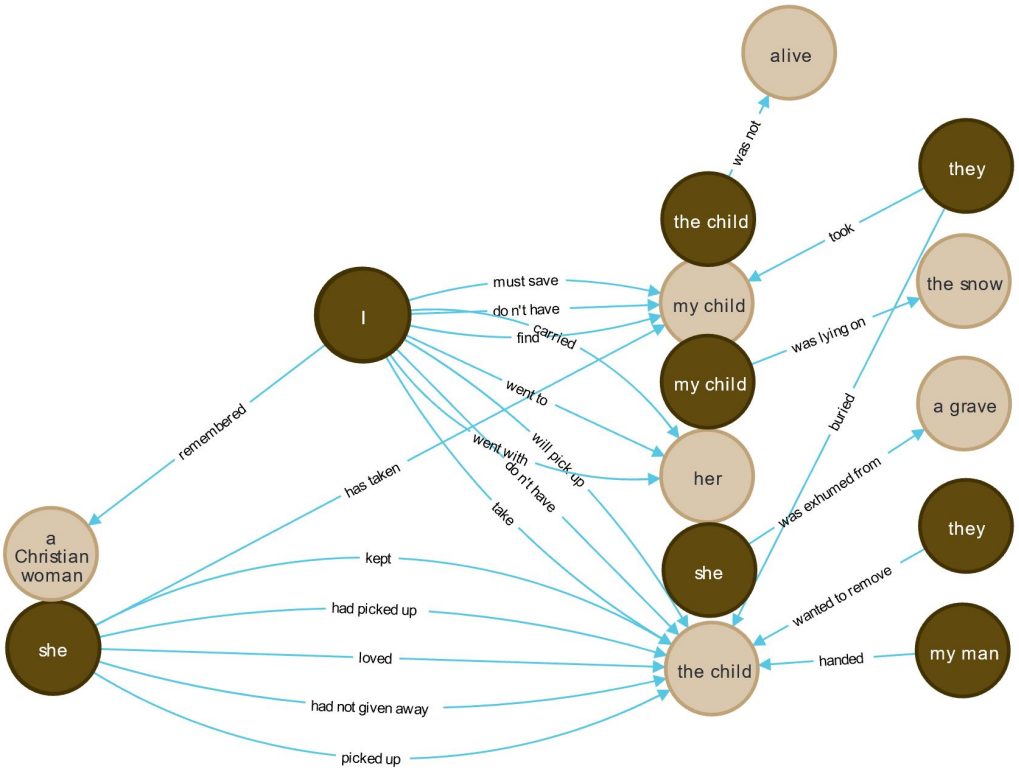


Figure 7. Building a semantic network from triplets related to the term *child* in Anna Kovitzka's testimony

4.1.2 Example 2: Erika Jacoby

After being deported to Auschwitz and separated from her grandparents on the arrival ramp, Jacoby and her mother are selected for the slave labor camp. Their heads are shaven; they are deloused for lice; and they are given meager clothes. Most of Jacoby's testimony, then, focuses on everyday life in Auschwitz, including forced labor, sleeping conditions, food rations, violence, and abuse. While the associated triplets bear witness to her struggle for survival, our indexing of agency in Jacoby's testimony also included the following groups of triplets. As shown in Table 3, after one passive construction, we see five successive, active constructions and one triplet characterized as *orientation* (considered as information to the interviewer). 32

Subjects	Relations	Objects	AEO Category
We	were	taken to a bath	Agency (Passive Speech)
we	were going by	a German officer's area	Agency (Active Speech)
I	saw	a swimming pool	Agency (Active Speech)
I	jumped out of	the line	Agency (Active Speech)
I	dove into	the swimming pool	Agency (Active Speech)
I	swam across	the pool	Agency (Active Speech)
They	didn't shoot	me	Orientation

Table 3. Erika Jacoby Triplets

This utterly astonishing action is not marked up or identified in any way through the Shoah Foundation's topical indexing. In fact, there is no word search that will find this part of her testimony. Jacoby relays the story in just a dozen sentences, comprising about 50 seconds of her two-hour testimony. And yet it is a stunning act of defiance, risk, and agency. 33

Here is the portion of the testimony with a little bit of context to set up the scene and reflect on its meaning: 34

> We stayed six weeks in Auschwitz. We had many, many interesting events that happened to us in Auschwitz. I just want to tell you one that was very important. And I don't think I have time to describe Auschwitz. And I'm sure that it's well known. But we were– one, one day, we were taken to a bath. We were taken to a bath about once a week or once every other week. And as we were marching, we were go– we were going by a German officer's area. And I saw a swimming pool. And I jumped out of the line, and I dove into the swimming pool. And I swam across the pool, and I got back. And they didn't shoot me, and I survived. But of course, my mother almost died. I mean, it was such an irr– irresponsible act from my part. But I was young, and I, I needed to be alive. And I did that. [Jacoby 1994, 51:40–52:30]

Indeed, there were two "pools" on the grounds of Auschwitz: one for prisoners and one for SS officers, near blocks 7 and 8. They were built as reservoirs to store water rather than be used for recreational swimming. Today, a sign clarifies the remains of the structure: "Fire brigade reservoir built in the form of a swimming pool, probably in early 1944." When Jacoby arrived in Auschwitz in the summer of 1944, the pool would have been complete. The Shoah Foundation manually tagged these segments of Jacoby's testimony with the following keywords: *camp intake procedures*, *camp prisoner marking*, *camp family interactions*, *loved ones' contacts*, *prisoner external contact*, and a single person tag, her mother, Malvina Salamonovits. Judging by these tags, one might expect her discussion to focus on procedural elements of the camp, perhaps coupled with how she and her mother survived together. There is nothing in the indexing that would indicate jumping out line, diving into the pool, and swimming across. While we can certainly read or hear Jacoby talking about this experience if we happened to land on this part of her testimony, it remains *an unindexed action*, which otherwise submerges a significant act of resistance and agency. Through the computational extraction of triplets, we can index, hone in on, and appreciate this defiant act. 35

After Auschwitz is liberated by the Soviet army, Jacoby relays that she and her mother left the camp in search of food and safety from sexual assaults by Russian soldiers. Over the next three and a half minutes (1:18-1:21), she describes a set of actions that helped her and her mother survive. She also tells of acts of revenge and anger committed by her 16-year-old self. Two of the three segments are tagged by the Shoah Foundation as *food acquisition* and *looting* (the third is not tagged at all). These topical terms convey little about the agency involved and the kinds of actions that a listener may want to search for and find in her testimony. In Table 4, we reproduce the full set of 23 semantic triplets that our methodology identified over these three minutes. All forms of agency are included as well as all subjects and objects. The "we" subject refers to Erika and her mother. 36

While *food acquisition* and *looting* are indexing terms that signal the overall set of content, these terms tell us nothing specific about what Erika and her mother did in the days after liberation; they tell us nothing about the vectors of agency and reasons for their actions; and they tell us none of the details of the actions conveyed in her testimony. And perhaps more pointedly, the indexing terms tend to obscure more than they reveal: Far from abstract topics, *food acquisition* involved the specific act of stealing and eating a pig (a non-kosher act of desperation for survival), while *looting* was connected to Jacoby's immediate post-War actions, which reflect personal anger, acts of revenge, and the longing for a normal life. The objects she takes from a home are symbols of family life, domesticity, and normalcy: a white tablecloth, an apron, and a little silver cup. Her act of *looting* is to take these everyday, material objects from somebody's home, the first possessions that she acquires after liberation from Auschwitz. 37

Subjects	Relations	Objects	Segments
we	got out of	the camp	1:18
we	went into	town	1:18
we	looked for	a house	1:18
we	found	a very beautiful home	1:18
we	occupied	it	1:18
we	couldn’t lock	the door	1:19
we	barricade	it	1:19
they	broke into	the stores	1:19
I	shoved	a lot	1:19
we	had	a canvas bag	1:19
we	passed by	a butcher store	1:19
I	took off	half a pig	1:19
I	carried	it	1:19
my mother	saw	it	1:19
she	cooked	the pig	1:19
we	ate	it	1:20
we	got	sick	1:20
I	was trying to find	a way	1:20
We	stayed in	this town	1:20
I	want to mention to	you	1:20
I	didn’t bring	it	1:20
I	have	a memento	1:20
I	broke into	a house	1:20
I	had	so much anger	1:20
I	expressed	that anger	1:20
I	went into	that house	1:20
I	broke	the piano	1:20
many people	went into	homes	1:21
I	did not want	anything	1:21
I	took from	the house	1:21
I	took	a white tablecloth	1:21
I	took	an apron	1:21
I	took	a little silver cup	1:21
it	had	the initials	1:21
I	kept	it	1:21
I	longed to establish	a normal life	1:21

Table 4. Erika Jacoby Triplets (1:18-1:21)

5 Discussion, Limitations, and New Work

Although our process of extracting and characterizing triplets creates new paratexts that enhance our interpretation and ability to index testimonies by descriptions of agency, there are a number of limitations and shortcomings. The major

issue is that “semantic triplets” – strictly speaking as subjects, verb relations, and objects – can oversimplify narrative expressivity when truncated, extracted, or decontextualized. Below, we will identify and discuss a number of limitations as well as discuss how our newer, “meticulous” triplet extraction process (version 2.0) addresses some of those problems.

5.1 Considering Coreferences: Pronoun Disambiguation and Entity Linking

As is evident in many of the triplets, pronouns often exist in subjects and objects, and they are sometimes ambiguous. For example, there are multiple references to “he,” “she,” “they,” or “them” where the pronoun could refer to a perpetrator in one line but a victim in the following. Thus, a coreferencing solution for the triplets needs to disambiguate pronouns or, at the very least, link triplets together such that the coreference can be clearly identified. We experimented with using NeuralCoref, a coreference resolution python package which annotates and resolves coreference clusters using a neural network [Hugging Face 2019]. However, this pre-trained language model is based on out-of-domain text data, and it is only effective about half the time, which is not precise enough for our needs. 39

Thus, we are working on creating our own model of coreferencing for use in Holocaust and genocide testimonies. This close-domain model was initially trained by annotating testimonies using BRAT, a web-based annotation tool that allows customized labels for entities and coreferences [Stenetorp et al. 2012]. Inspired by the annotation process of entities and coreferences in an English literature dataset, we will use a customized set of entity types, including named people, groups, communities, as well as other regular named-entities like organizations and geo-political entities [Bamman et al. 2019]. Our hope is that the training model, derived from a diverse set of annotated Holocaust testimonies, will help us disambiguate pronouns throughout a corpus. 40

5.2 Revisiting the Chunk Method: A More “Meticulous” Triplet Extraction Process

Our chunk-based triplets extraction method is certainly not flawless. One challenge is that important expressions that are not triplets are simply not found (for instance, “we were separated”). Another challenge is how to capture fuller object phrases and distinguish direct and indirect objects. In this version, we capture both, and usually the direct object is part of the triplet and the indirect object is part of the contextual information; in the more “meticulous” version, we capture fuller object phrases by extracting from the “context” as much as possible. Although the objects tend to be longer, they are more complete and integrate the contextual elements within the object phrase itself. This newer version results in both more triplets being extracted and higher accuracy.[15] 41

An ongoing challenge, however, concerns certain sentences with dependent, relative, or conditional clauses, which are not always captured completely or may be broken apart in the course of the chunk method. This problem can be seen with “if, then” statements or ones that begin with a form of negation. For example, this sentence spoken by Bassfreund, is turned into three triplets: “And it was already night and the SS opened the doors and said if we throw out the dead bodies we shall get some food.”[Bassfreund 1946] Getting food from the SS is dependent upon the victims being forced to throw out dead bodies from the train. Although the triplets are captured individually, maintaining the specific contingency of the actions and understanding the reasons for them requires human review and interpretation. 42

While network analysis may give rise to new narrative orderings that allow us to see relationships across a testimony or corpus, we also preserve the narrative sequence of the triplets in our extraction process since the context for a triplet certainly matters both within an individual sentence and between sentences. For instance, when Bass says: “Well, you could tell. These guys were very rough and very bad. They used to beat you up, and scream, and holler, and push you. And hit you with a– a rubber– piece of rubber there. It looked like a rubber hose.”[Bass 1997] Our triplet extraction method maintains the contiguity between the description of the men (former criminals who the Nazis put in place as supervisors in camps) so that the last triplet (“it looked like a rubber hose”) continues to make sense within this context. Sequential groupings are one way to maintain narrative continuity. 43

Another basic challenge that we face is extracting triplets from fragmented, vernacular speech, which the process often has trouble parsing. Survivors sometimes switch from first-person narration to second or third-person; sometimes the sentences contain incomplete thoughts, repetitions, or stuttering; sometimes the transcribed sentence lacks regular 44

punctuation; and sometimes survivors quote the speech of others directly or indirectly. The triplet extraction process has no way of telling if a triplet refers to the quoted or attributed speech of someone else. For example, Anna Kovitzka says: “Once my man said: ‘I can’t make peace with them. Our child must be saved even if we two shall die.’” [Kovitzka 1946] While two triplets are extracted (“I can’t make peace with them” and “our child must be saved”), the triplets are not attributed to her husband because the phrase “once my man said” is not part of either triplet.

To take another example: In his interview with Boder, Bassfreund says: “We were rubbed with kerosene” [Bassfreund 1946] with respect to the delousing process in Auschwitz. That triplet is easily found; however, in his later interview with the Shoah Foundation, he says: “And then they did some – used some kind of delousing agent, they poured all over us.” No triplets were found in the first version of our process, as the sentence is spoken with breaks that less clearly connect the subject (“they”) to the verbs and objects. To address this, our “meticulous” version of triplet extraction is based on a set of finer-defined, semantic rules to extract noun and verb chunks. To make chunks include more contextual information and complete grammatical structures, we add a layer in between the sentence and chunk levels, which we call the “segment level,” for chunk-boundary decision making, as shown in Figure 8. Even though the sentence is not grammatically correct, the new extraction method is able to find the triplets in this sentence, while the previous approach was not. And although some objects and object phrases become longer than expected, we can now extract a more complete context for the objects of a triplet. We are also experimenting with categorizing object phrases based on semantic features like word lemmas, WordNet synsets, and dependency parsing, which will create more filters for large scale triplets retrieval.

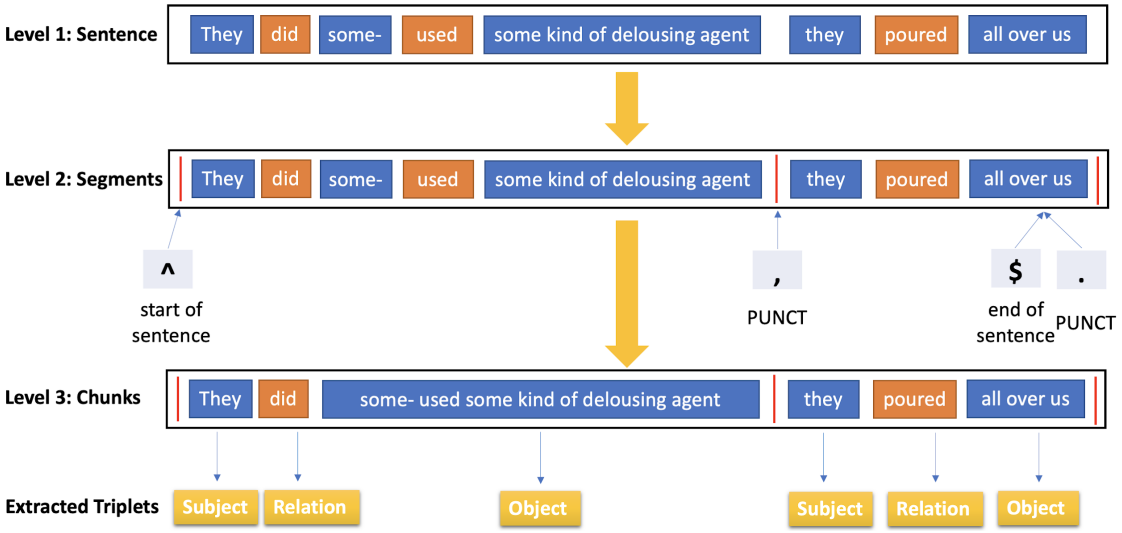


Figure 8. Example of the Three-level Chunk-based Triplets Extraction Method

5.3 Other Improvements and Potential Applications

Based on the indexing and analysis of agency in the four example testimonies discussed here, we have applied our “meticulous” method to over a thousand Holocaust testimonies. The output creates a new paratext of triplets for each interview and provides a set of additional metadata that can be queried, indexed, and overlaid on existing metadata. The triplets are intended to function as finding aids to supplement the original text by providing a new set of metadata attuned to articulations of agency. One of the promising directions is to query “shared experiences” (such as reports of and responses to discrimination) across a corpus using the semantic triplets. What, for example, did hundreds of different witnesses say about their memories of discrimination during the 1930s and early 40s and how did they describe their individual reactions? How many other people remember removing the yellow star from their clothes and what did they say they did next?

Going forward, we hope to create a better finding aid and scaled-up visualization tool, perhaps through an indexing dashboard supported by a triplets database. Researchers can then query and visualize through the same interface to

find semantic triplets and generate associated networks. This interactive dashboard could combine the flexibility of curation in the Triplets Dashboard in Tableau with a Neo4j Network Visualization Dashboard. At the same time, we are continuing to explore how searching for “agency” within a narrative can complement standard thematic and topical searches through network clustering and category clustering of triplets.

6 Conclusion

Semantic triplets are a text immanent way to computationally identify and index descriptions of agency within oral histories. While conventional indexing is important for finding overarching themes and topics, it relies on the creation of a secondary scaffolding for searching within the testimonies, almost always attuned to named entities, topics, and general themes. Indexing semantic triplets, on the other hand, is primarily an algorithmic approach to identify specific expressions of agency and their lexical context from transcripts. As such, we are able to link and analyze multiple vectors of action and attune our reading and listening to networks of agents, relations, objects, and contexts. Our approach always relies on and is derived from the interviewee’s spoken (and transcribed) language. In this regard, it is a kind of “algorithmic close reading” built on a paratext derived from the original transcript. Although “distant reading” remains a powerful approach for statistically summarizing the content of a corpus, digital humanities scholars have also warned against the seduction of “delegating observation” and “delegating interpretation” to machines [Underwood 2019, 157–158]. In many ways, semantic triplets use some of the algorithmic tools of distant reading (large-scale text analysis attuned to patterns, classifications, and clusters) to enable humans to undertake close readings and perform interpretative queries. It is up to us to figure out the meaning or significance of any description of agency in the context of an oral history or corpus. 48

In this sense, our approach is also quite different from current machine learning methods for information extraction systems. Although seemingly unpopular these days from a natural language processing perspective, a fully rule-based and easily interpretable system works well for our purposes of triplets extraction and analysis. We can describe the algorithms we use based on conditional statements and clear rules for additional human curation and correction. Computational indexing of agency shifts the scale of analysis to individual mentions of agency and interconnected actions, helping us read and listen to testimonies in ways that deepen our understanding of what people report they did and what was done to them. 49

7 Acknowledgements

The research presented here owes a significant debt of gratitude to the USC Shoah Foundation, particularly Stephen Smith, Samuel Gustman, Martha Stroud, Claudia Wiedeman, and Crispin Brooks. All of the analyses presented in this paper were performed at UCLA under the direction of Todd Presner. The co-authors, Lizhou Fan and Todd Presner, are grateful to UCLA team members Anna Bonazzi, Rachel Deblinger, Kyle Rosen, Michelle Lee, and Wanxin Xie. We also thank Anne Knowles, David Shepard, Anthony Caldwell, Wolf Gruner, and Zoe Borovsky for their generous support and feedback over the years. 50

Appendix

Algorithm 1 Triplets AEO Algorithm

Input: Spacy Tokens for a Relation R and Spacy Tokens for an Object O, Evaluation Verbs $list_{evaluation}$, Orientation Verbs $list_{orientation}$, Speculative Action Verbs $list_{posact}$
Output: Triplets AEO Category C_{AEO};

Step 1: Initialization
Initialize 0 Integers as Status Identifiers, including $r_{has\_evaluation}$, $r_{has\_orientation}$, r_{has_posact}, r_{has_be}, r_{has_have}, r_{has_to}, r_{has_neg}, r_{has_VBG}, r_{num_verb}, o_{is_adj}, and o_{has_no}
# Step 2: Value Assignments for R
for $r \in R$ **do**
    if lemma of $r \in list_{evaluation}$ **then** Assign 1 to $r_{has\_evaluation}$
    else if lemma of $r \in list_{orientation}$ **then** Assign 1 to $r_{has\_orientation}$
    else if lemma of $r \in list_{posact}$ **then** Assign 1 to r_{has_posact}
    else if lemma of r *is word be* **then** Assign 1 to r_{has_be}
    else if lemma of r *is word have* **then** Assign 1 to r_{has_have}
    else if lemma of r *is word to* **then** Assign 1 to r_{has_to}
    else if semantic dependency tree tagger of r *is label neg* **then** Assign 1 to r_{has_neg}
    end if
end for
# Step 3: Value Assignments for O
for $o \in O$ **do**
    if lemma of o *is word no* **then** Assign 1 to o_{has_no}
    end if
end for
for $o \in O$ **do**
    if part of speech tagger of o *is label* ADJ **then** Assign 1 to o_{is_adj}
    end if
    if part of speech tagger of $o \in$ *labels* $NOUN, PROPN, PRON$ **then** Assign 0 to o_{is_adj} and end For loop
    end if
end for
# Step 4: AEO Category Decision
if $r_{has\_evaluation}$ and o_{is_adj} **then** C_{AEO} = Evaluation
else if r_{has_posact} **then** C_{AEO} = Agency_Possible
else if $r_{has\_orientation}$ **then** C_{AEO} = Orientation
else if r_{has_neg} or o_{has_no} **then** C_{AEO} = Orientation
else if r_{has_have} **then**
    if r_{has_to} **then** C_{AEO} = Agency_Coercive
    else C_{AEO} = Orientation
    end if
else if r_{has_be} **then**
    if o_{is_adj} **then** C_{AEO} = Evaluation
    else if r_{has_VBG} **then** C_{AEO} = Agency_Active
    else if $r_{num\_verb} > 1$ **then** C_{AEO} = Agency_Passive
    else if $r_{num\_verb} = 1$ **then** C_{AEO} = Orientation
    end if
else C_{AEO} = Agency_Active
end if

Figure 9. Triplets AEO Algorithm, as an Example of High Interpretability

Notes

[1] Removing the “yellow patch” refers to removing the yellow star that Jews were forced to wear on their outer garments to identify themselves publicly.

[2] Both before and after the time that Boder completed his audio recordings, various Historical Commissions throughout Germany, Poland, Hungary, and Eastern Europe also collected personal stories, primarily through written questionnaires and interviews [Jockusch 2012].

[3] The best biography of Boder and the significance of his recordings is *The Wonder of their Voices: The 1946 Holocaust Interviews of David Boder* [Rosen 2010].

[4] Boder and his interviewees do not use the term “Holocaust,” as this term did not come into common usage until the 1950s to describe the Nazi genocide of the Jews. He does, however, refer to and attempt to explicate “the concentration camp phenomenon” (xiv) in the preface to his book of interviews, *I Did Not Interview the Dead* (1949). He says that the “displaced” and “uprooted” people were “dislocated by a world catastrophe” (xviii), perhaps referencing the Hebrew word *Shoah* or the Yiddish word *Churban*, both of which were used by his interviewees to describe the destruction of the Jewish communities of Europe.

[5] Copies of the original wire spools were sent to the Library of Congress, which, decades later, procured the technical means to undertake the media transfer of the wire-recorded audio to tape. Later, they were digitized by the Illinois Institute of Technology (IIT) and made available as WAV files through the Aviary platform. The interviews can be accessed on the Voices of the Holocaust website: https://voices.library.iit.edu/david_boder. The English translations, annotations, and indexing were done by Boder for the testimonies in his book, *Topical Autobiographies of Displaced People* [Boder 1950-56].

[6] Boder's full indices are found at the end of *Topical Autobiographies of Displaced People* [Boder 1950-56, 3105–3159].

[7] The Shoah Foundation generously provided our team with 983 English-language transcripts prepared by ProQuest and an additional set of 900 German language testimonies prepared by the Freie University in Berlin. While this article describes only the analyses conducted with English-language transcripts, the methods are being developed for German, too.

[8] For simplicity and consistency, we still use the name *Noun Chunk* to represent the extended version, i.e. nouns and/or adjectives (NAP).

[9] Our work is built on the pretrained SpaCy model "en_core_web_lg," which is an English pipeline with components including part of speech tagger, dependency parser, and lemmatizer.

[10] Our method is inspired by the built-in SpaCy function of noun chunk extraction, as described in its python source code syntax_iterators.py.

[11] As discussed in section 5, we have also developed a fuller, chunk-based extraction method that enables us to capture longer object phrases with higher accuracy.

[12] The process of assigning "victim" and "perpetrator" identities is very incomplete due to the need to disambiguate pronouns ("they," "he," "she") as well as account for proper names.

[13] The German original also shows both passive and coerced speech: "Bevor wir abtransportiert wurden, mussten wir unterschreiben, dass wir wegen staatspolitischer und feindlicher Umtriebe Deutschland verlassen müssen." [Bassfreund 1946]

[14] Examples in the original German, which all follow the semantic pattern of passive speech in the English translations, include: "Plötzlich waren wir umringt von der Leibstandarte, und wurden dann mit Füßen getreten und in Autos verladen"; "Und dann kamen wieder Autos und wieder wurden wir von der Leibstandarte zu einem ganz entlegenen Bahnhof in Berlin gebracht und wurden dort einwagoniert"; "wir wurden aus diesen Wagons herausgetrieben"; "wir wurden mit Petroleum eingerieben." [Bassfreund 1946]

[15] The details of the "meticulous" version, including the rules and sample data, are available in the Github repository: https://github.com/lizhouf/semantic_triplets

Works Cited

Abney 1996 Abney, Steven. 1996. "Partial Parsing via Finite-State Cascades." *Nat. Lang. Eng.* 2 (4): 337–344. https://doi.org/10.1017/S1351324997001599.

Angeli et al. 2015 Angeli, Gabor, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. "Leveraging Linguistic Structure For Open Domain Information Extraction." In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 344–354. Beijing, China: Association for Computational Linguistics. https://aclanthology.org/P15-1034.

Apache Software Foundation 2017 The Apache Software Foundation. 2017. "Welcome to Apache OpenNLP." 2017. http://opennlp.apache.org/.

Bamman et al. 2019 Bamman, David, Olivia Lewke, and Anya Mansoor. 2019. "An annotated dataset of coreference in English literature." *arXiv preprint arXiv:1912.01140*.

Bass 1997 Bass, Jack. Interview 30765. *Visual History Archive*, USC Shoah Foundation, 1997. Accessed November 1, 2020.

Bassfreund 1946 Bassfreund, Jürgen. Interview with David Boder (September 20, 1946). *Topical Autobiographies of Displaced People*. 276-318. Also available on: *Voices of the Holocaust*, Illinois Institute of Technology. https://voices.library.iit.edu/interview/bassfreundJ. Accessed November 1, 2020.

Boder 1949 Boder, David. 1949. *I Did Not Interview the Dead*. University of Illinois Press.

Boder 1950-56 Boder, David. 1950-56. *Topical Autobiographies of Displaced People*. UCLA Special Collections, Young Research Library. Unpublished manuscript. Boxes 9-11.

Bradley and Pasin 2017 Bradley, John, and Michele Pasin. 2017. "Fitting Personal Interpretation with the Semantic Web: Lessons Learned from Pliny." *DHQ: Digital Humanities Quarterly* 11 (1).

Explosion AI 2020 Explosion AI. 2020. "Industrial-Strength Natural Language Processing." https://spacy.io/.

Fader et al. 2011 Fader, Anthony, Stephen Soderland, and Oren Etzioni. 2011. “Identifying Relations for Open Information Extraction.” In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1535–1545. Edinburgh, Scotland, UK.: Association for Computational Linguistics. https://www.aclweb.org/anthology/D11-1142.

Gruner 2011 Gruner, Wolf. 2011. “‘The Germans Should Expel the Foreigner Hitler’: Open Protest and Other Forms of Jewish Defiance in Nazi Germany.” Edited by David Silberklang. *Yad Vashem Studies* 39 (2): 13–53.

Gruner 2016a Gruner, Wolf. 2016a. “Defiance and Protest. A Comparative Micro-Historical Re-Evaluation of Individual Jewish Responses towards Nazi Persecution.” Edited by Claire Zalc and Tal Bruttmann. *Microhistories of the Holocaust*, 209–26.

Gruner 2016b Gruner, Wolf. 2016b. “Defiance and Protest: Forgotten Acts of Individual Jewish Resistance.” https://www.cornell.edu/video/wolf-gruner-jewish-defiance-protest-nazi-germany.

Hilberg 1992 Hilberg, Raul. 1992. *Perpetrators, Victims, Bystanders: The Jewish Catastrophe, 1933-1945*. Aaron Asher Books New York.

Hugging Face 2019 Hugging Face. 2019. “NeuralCoref 4.0: Coreference Resolution in SpaCy with Neural Networks.” 2019. https://github.com/huggingface/neuralcoref.

Hyvönen 2020 Hyvönen, Eero. 2020. “Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-Analysis and Serendipitous Knowledge Discovery.” *Semantic Web*,11(1): 187-193.

Jacoby 1994 Jacoby, Erika. Interview 8. Segments, 52-53, 78-82. *Visual History Archive*, USC Shoah Foundation, 1994. Accessed November 1, 2020.

Jockers and Underwood 2015 Jockers, Matthew L., and Ted Underwood. 2015. “Text-Mining the Humanities.” *A New Companion to Digital Humanities*, Edited by Susan Schreibman, Ray Siemens, and John Unsworth. John Wiley and Sons. 291–306.

Jockusch 2012 Jockusch, Laura. 2012. *Collect and Record! Jewish Holocaust Documentation in Early Postwar Europe*. Oxford University Press.

Kovitzka 1946 Kovitzka (Kaletska), Anna. Interview with David Boder (September 26, 1946). *Topical Autobiographies of Displaced People*. 244-275. Also available on: *Voices of the Holocaust*, Illinois Institute of Technology. https://voices.library.iit.edu/interview/kaletskaA. Accessed November 1, 2020.

Labov and Waletsky 1997 Labov, William, and Joshua Waletzky. 1997. “Narrative Analysis: Oral Versions of Personal Experience.” *Journal of Narrative & Life History* 7 (1–4): 3–38.

Luft 2015 Luft, Aliza. 2015. “Toward a Dynamic Theory of Action at the Micro Level of Genocide: Killing, Desistance, and Saving in 1994 Rwanda.” *Sociological Theory* 33.2: 148–72.

Mausam et al. 2012 Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. 2012. “Open Language Learning for Information Extraction.” In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 523–534. Jeju Island, Korea: Association for Computational Linguistics. https://www.aclweb.org/anthology/D12-1048.

Peter 2001 Peter, Dingley Andrew. 2001. Data replication system and method. US6304882B1, issued October 16, 2001. https://worldwide.espacenet.com/patent/search/family/009930067/publication/US2003145022A1?q=US2003145022.

Presner 2016 Presner, Todd. 2016. “The Ethics of the Algorithm: Close and Distant Listening to the Shoah Foundation Visual History Archive.” In *Probing the Ethics of Holocaust Culture*, edited by Claudio Fogu, Wulf Kansteiner, and Todd Presner, 175–202. Harvard University Press.

Rosen 2010 Rosen, Alan. 2010. *The Wonder of Their Voices: The 1946 Holocaust Interviews of David Boder*. Oxford University Press.

Rothberg 2019 Rothberg, Michael. 2019. *The Implicated Subject: Beyond Victims and Perpetrators*. Stanford University Press.

Rusher 2006 Rusher, Jack. 2006. “Triple Store.” 2006. https://www.w3.org/2001/sw/Europe/events/20031113-storage/positions/rusher.html.

Saldias and Roy 2020 Saldias, Belen, and Deb Roy. 2020. “Exploring Aspects of Similarity between Spoken Personal Narratives by Disentangling Them into Narrative Clause Types.” In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, 78–86. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.nuse-1.10.

Stenetorp et al. 2012 Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. "BRAT: a web-based tool for NLP-assisted text annotation." In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 102-107. 2012.

Swanson et al. 2014 Swanson, Reid, Elahe Rahimtoroghi, Thomas Corcoran, and Marilyn Walker. 2014. "Identifying Narrative Clause Types in Personal Stories." In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 171–180. Philadelphia, PA, U.S.A.: Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-4323.

USC Shoah Foundation 2006 USC Shoah Foundation. 2006. *Indexing Guidelines*. https://sfi.usc.edu/sites/default/files/docfiles/Indexing_Guidelines_0.pdf.

Underwood 2019 Underwood, T. 2019. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.

Voices of the Holocaust *Voices of the Holocaust*. Illinois Institute of Technology, Paul V. Galvin Library. https://voices.library.iit.edu/.

Wieviorka 2006 Wieviorka, Annette. 2006. *The Era of the Witness*. Translated by Jared Stark. Cornell University Press.

Wu and Weld 2010 Wu, Fei, and Daniel S. Weld. 2010. "Open Information Extraction Using Wikipedia." In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 118–127. Uppsala, Sweden: Association for Computational Linguistics. https://www.aclweb.org/anthology/P10-1013.

Yates et al. 2007 Yates, Alexander, Michele Banko, Matthew Broadhead, Michael Cafarella, Oren Etzioni, and Stephen Soderland. 2007. "TextRunner: Open Information Extraction on the Web." In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 25–26. Rochester, New York, USA: Association for Computational Linguistics. https://www.aclweb.org/anthology/N07-4013.

Zalc and Bruttman 2016 Zalc, Claire, and Tal Bruttmann, eds. 2016. *Microhistories of the Holocaust*. Berghahn Books.