The N2 corpus: A semantically annotated collection of Islamist extremist stories

Mark A. Finlayson[1], **Jeffry R. Halverson**[2], **Steven R. Corman**[3]

[1]Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
32 Vassar Street, Room 32-258, Cambridge, MA 02139
markaf@mit.edu

[2]Department of Philosophy and Religious Studies
Coastal Carolina University
PO Box 261954, Conway SC 29528
halverso@coastal.edu

[3]Center for Strategic Communication
Arizona State University
Box 871205, Tempe, AZ 85287
steve.corman@asu.edu

Abstract

We describe the N2 (Narrative Networks) Corpus, a new language resource. The corpus is unique in three important ways. First, every text in the corpus is a story, which is in contrast to other language resources that may contain stories or story-like texts, but are not specifically curated to contain only stories. Second, the unifying theme of the corpus is material relevant to Islamist Extremists, having been produced by or often referenced by them. Third, every text in the corpus has been annotated for 14 layers of syntax and semantics, including: referring expressions and co-reference; events, time expressions, and temporal relationships; semantic roles; and word senses. In cases where analyzers were not available to do high-quality automatic annotations, layers were manually double-annotated and adjudicated by trained annotators. The corpus comprises 100 texts and 42,480 words. Most of the texts were originally in Arabic but all are provided in English translation. We explain the motivation for constructing the corpus, the process for selecting the texts, the detailed contents of the corpus itself, the rationale behind the choice of annotation layers, and the annotation procedure.
Keywords: narrative corpora; multi-layered annotation; religious texts

We describe the N2 (Narrative Networks) Corpus, a new language resource. The corpus is unique in three important ways. First, every text in the corpus is a story, which is in contrast to other language resources that may contain stories or story-like texts, but are not specifically curated to contain only stories. Second, the unifying theme of the corpus is material that Islamist Extremists have produced, or is often referenced by them. The texts in the corpus include: personal narratives gathered from internet forums; press releases describing bombings and attacks by extremist groups in Afghanistan; articles containing stories included in al-Qaeda propaganda materials (the Inspire magazine); and religious stories (Hadith and Sirah) often referenced by extremist groups. Third, every text in the corpus has been annotated for 14 layers of syntax and semantics: tokens; sentences; part of speech tags; lemmas; context-free grammar parses; referring expressions; co-reference groups; events; time expressions; temporal relationships; referent properties; multi-word expressions; semantic roles; and word senses. The corpus was annotated with a mix of automatic, semi-automatic, and manual methods. The corpus comprises 100 texts and 42,480 words, most originally in Arabic but all translated into English. We present here a comprehensive view of the corpus: our motivation, the process for selecting the texts, the detailed contents, the rationale behind the choice of annotation layers, and the annotation procedure.

1. Motivation

We had two motives for creating the N2 Corpus. First, radical religion extremism is an acute and growing problem, and yet the dynamics of the radicalization of marginal populations is not well understood. Furthermore, it is has been suspected for some time that stories have a powerful effect on the radicalization process (Halverson et al., 2011). Therefore, we sought to provide a resource for those studying religious radicalization via stories.

Second, computational modeling of narrative has of late been a burgeoning field of research (Finlayson et al., 2013). Work in this area, however, suffers from a lack of comprehensive annotated data that provides the sort of basic *who does what to whom* information necessary for story understanding. There have been recent proposals to create a "Story Bank" comprised of a "handful of handfuls" (Finlayson, 2011a). Our goal was provide one high-quality handful targeted at a specific problem (understanding religious radicalization via stories). If one's measure combines both the number of texts along with their depth of annotation, N2 Corpus is now the most extensive annotated corpus of stories available.

2. Corpus contents & selection critera

The corpus can be split into three distinct sections: Hadith and Sirah; OpenSource.gov Extremist Texts; and Inspire

Text Type	Text Source	Number of Texts	Number of Words
Current Affairs	OpenSource.gov	13	5,419
	Inpire	7	7,814
History	OpenSource.gov	8	2,480
Biography	OpenSource.gov	8	11,795
Religious Texts	Hadith & Sirah	64	14,972
	Total	**100**	**42,480**

Table 1: Text types, sources, and counts.

magazine stories. The corpus contains 100 texts totaling 42,480 words. A detailed breakdown of the corpus contents is shown in Table 1.

2.1. Hadith and Sirah

The first section contains religious texts widely read and studied within Islam, specifically from the Sunni tradition, that are often selectively used by Islamist extremists. These texts included the major Hadith collections (reports of the deeds and sayings of Muhammad), such as al-Bukhari (Abu Dawud, 2000), and the *Sirah* or hagiography (sacred biography) of the Prophet Muhammad by Ibn Hisham (which is itself an edited version of an earlier *Sirah* by Ibn Ishaq) (Ibn Hisham, 2000). The Hadith are collections of thousands of individual oral traditions (later written down) relating different sayings and events attributed to the Prophet Muhammad and his companions (*sahaba*). Sunni Muslims recognize six different Hadith collections as trustworthy or "canonical" alongside other influential collections, such as the *Muwatta* of Imam Malik (Imam Malik, 2005). The set of sources drawn on are listed in Table 2.

The majority of the content in the Hadith and Sirah have no direct connection to Islamist extremists. We specifically extracted only material discussing warfare (military *jihad*), martyrdom, infidels (*kafirun*), corporal punishments (e.g. stoning adulterers), and the status of apostates and blasphemers, for inclusion in the corpus. This material has special relevance to Islamist extremists who use it to justify acts of violence and delineate the "true" community of Muslims from the majority of Muslims in the world. This section of the N2 Corpus contains 64 texts totalling 14,972 words. All of these texts were originally in Arabic, and translated into English within the last two decades.

Type	Collection	Reference
Hadith	Malik's Muwatta	(Imam Malik, 2005)
Hadith	Sunan Abu Dawud	(Abu Dawud, 2000)
Hadith	Sahih al-Bukhari	(al Bukhari, 1994)
Hadith	Sahih Muslim	(Muslim ibn al Hajjaj, 2000)
Sirah	Sirat Ibn Hisham	(Ibn Hisham, 2000)

Table 2: Sources of Hadith & Sirah in the N2 Corpus.

2.2. OpenSource.gov Extremist texts

Islamist extremist texts in the corpus were drawn from a database containing a curated collection of public statements, video transcripts, and blog/forum posts extracted primarily from the open source intelligence portal OpenSource.gov. This resource contains unclassified material collected and translated into English by the United States government. It is available to researchers and others working under U.S. government funding, including academics. Items in the collection were collected by trained curators and linguists who selected texts that (1) were distributed by al-Qaeda and affiliated groups, (2) pertained to conflicts in the Middle East, North Africa, and Southeast Asia, and (3) were likely to contain stories. Once collected, paragraphs of the texts were coded into categories that, among other things, distinguished story paragraphs from non-story paragraphs using a coding system established as reliable (Ceran et al., 2012). This section of the corpus contains 29 texts totaling 19,694 words. Many of these texts were originally in Arabic or another local language (e.g., Urdu), but were translated into English.

2.3. Inspire magazine

Inspire magazine is a publication of al-Qaeda in the Arabian Peninsula (AQAP) professionally produced in the English language and widely distributed via social media channels. Conceived by the late American extremist Anwar al-Awlaki, it contains articles about Islamist extremist ideology and methods for conducting terrorist operations. It is intended to influence potential extremists in Western countries and to bolster al-Qaeda's "brand." Eleven issues have been published to date, from July 2010 to June 2013. Many analysts believe that Inspire has been influential in promoting extremist ideology and methods. We selected from the available magazines a selection of articles that were either actual stories (there was a running column called "Jihad Stories") or story-like in form. In particular, we included the narrative-like introduction to the article "How to Make a Bomb in the Kitchen of Your Mom", which purportedly provided the instructions used in making the pressure-cooker bombs exploded at the Boston Marathon on April 15, 2013. This section of the corpus contains 7 texts totaling 7,814 words. Many of these texts were originally written in English; some seem to be translated from interviews with the story tellers.

3. Annotation layers

The main difficulty in assembling corpora useful to the computational modeling of narrative is not finding the stories themselves. It is, rather, producing a computer-readable interpretation of the information in the story useful for the modeling task at hand. To address this problem, the 100 texts of the N2 corpus were annotated with 14 layers of

#	Group	Layer	Annotation Style	Strict F_1	Layer-Specific Measures
1	Syntax	Tokens	Automatic, w/ corr.	-	-
2		Sentences	Automatic, w/ corr.	-	-
3		Part of Speech Tags	Semi-Automatic	0.97	0.95 (category κ)
4		Lemmas (stems)	Semi-Automatic	0.95	-
5		CFG Parses	Automatic	-	-
6	Referential Structure	Referring Expressions	Manual	0.89	0.94 (loose F_1)
7		Coreference Groups	Manual	0.65	0.81 (chance-adjusted Rand)
8	Temporal Structure	Time Expressions (TimeML)	Manual	0.40†	0.69 (ident. F_1), 0.77 (cat. κ)
9		Events (TimeML)	Manual	0.58†	0.82 (ident. F_1), 0.73 (cat. κ)
10		Time Links (TimeML)	Manual	0.52†	-
11	Word Semantics	Referent Properties	Manual	0.87	0.92 (ident. F_1), 0.91 (cat. κ)
12		Multi-word Expressions	Semi-Automatic	0.65	-
13		Word Senses	Manual	0.72	-
14		Semantic Roles	Semi-Automatic	0.40†	0.59† (core args F_1)

Table 3: Layers of Annotation. Agreement is reported both as a strict F_1 measure, which measures exact syntactic agreement between annotators, as well as in layer-specific measures. Strict F_1 generally under-estimates agreement significantly. Low measures are marked with † and are discussed in more detail in Section 3.

syntax and semantics which are roughtly sufficient to express the *who does what to whom* of the stories. These layers, plus their detailed agreement measures, are shown in Table 3. The layers may be broken down into four different groups: Syntax, Referential Structure, Temporal Structure, and Word Semantics.

3.1. Annotation style & agreement

Layers were annotated in several different ways, as indicated in Table 3. **Automatic** means the layer was calculated automatically and not corrected or adjusted by hand. **Automatic, with corrections** means that the layer was automatically calculated, and was corrected unilaterally by the project manager when an error was discovered. **Semi-Automatic** means the layer was first annotated automatically, and then these annotations were hand-corrected by annotators in a double-blind, adjudicated procedure. **Manual** means the layer was annotated completely from scratch by hand in a double-blind, adjudicated procedure.

We report several different types of agreement measures. First, we report a uniform measure across all layers, which is the strict syntactic F_1 measure, which is calculated by examining when the annotators for double-annotated layers agreed exactly. This measure significantly under-estimates actual agreement on many aspects of each annotation task, especially in the case of complex or multi-featured annotations such as semantic roles and time links. Layer-specific measures are discussed in the following sections.

3.2. Syntax

The first group contains the syntactic layers. These layers were mostly calculated automatically. In particular, tokens and sentences were calculated using the Stanford CoreNLP Tokenizer and Sentence Detector. These were not subject to any additional annotation; if an error was discovered this was corrected unilaterally by the project manager.

Lemmas were calculated automatically using the JWI implementation of the Wordnet Morphy stemmer (Finlayson, 2014). Part of speech tags were calculated using the Stanford POS tagger (Toutanova et al., 2003). As a layer-specific measure, we report the Fleiss κ for assigning POS tags. This measure was excellent, at 0.95. Both POS tags and lemmas were corrected in the course of annotating word senses. Finally, the Stanford CFG grammar parser (Klein and Manning, 2003) was used to provide a CFG parse for every sentence. The CFG parses were not corrected in any way.

In general, the Stanford CoreNLP suite was used as a matter of convenience: because the Story Workbench annotation tool (discussed in the next section) is written in Java, it was easiest to use NLP tools also written in Java. The Stanford suite is also fast, relatively bug-free, and is known to have extremely good (if not state-of-art) performance.

3.3. Referential structure

Annotators manually marked the text for referring expression and co-reference relationships, as described in (Hervas and Finlayson, 2010). We defined referring expressions as referential noun phrases and their coreferential expressions, e.g., "John kissed Jane. She blushed." This included referring expressions to generics (e.g., "*Lions* are fierce."), dates, times, and numbers, as well as events if they were referred to using a noun phrase. We included in each referring expression all the determiners, quantifiers, adjectives, appositives, and prepositional phrases that syntactically attached to that expression. When referring expressions were nested, all the nested referring expressions were also marked separately. In addition to strict F_1 measures, we calculated two additional agreement measures: a "loose" F_1 for referring expressions, which, in addition to exact matches, identified matches for non-exactly matching referring expressions from overlapping candidates. This has the effect of ignoring minor inconsistencies in choice of referring expression boundaries (for example, a common mistake was forgetting to include an associated article such as *a*, *an*, or *the*). The second measure we calculated was the chance-adjusted Rand index (Hubert and Arabie, 1985), which was applied to the coreference groups. This measure has a range

from -1 (anti-correlated groups) to 0 (no correlation between groups) to 1 (exactly matching groups). The agreement here was quite good, at 0.81.

3.4. Event structure

We applied the TimeML suite of annotations to the corpus (Pustejovsky et al., 2003). This can be split into three layers: Events, Time Expressions, and Time Links. Because the automatic analyzers available for TimeML are not that good, this annotation was done completely manually by the annotators.

Events are defined as happenings or states, and can fall into one of seven different classes: Occurrence, Reporting, Perception, Aspectual, Intensional Action, State, or Intensional State. In addition to the main class of the event, annotators marked the event head, the event's full extent, the polarity and modality (plus any associated tokens), and a few syntactic features such as part of speech, tense, aspect, and person. While the strict F_1 measure was not that good, at 0.58, two additional measures indicated that the annotators actually were agreeing quite well, except for minor details. We calculated both an "identification" F_1 and the Fleiss κ for assignment of the class of the event. The identification measure was intended to capture how well annotators agreed that there was an event present, regardless of whethey identified the boundaries or various sub-features exactly the same. This was good, at 0.82. The Fleiss κ for the event class was 0.73, which indicates "very good" agreement. These agreement numbers would seem to indicate that the annotators usually agreed when an event was present, and would also usually agree on the event class; but that there was substantial variation regarding the other minor features of the markings.

Time expressions mark the location and type of temporal expressions. Each time expression is a sequence of tokens, potentially discontinuous, that indicate a time or date, how long something lasted, or how often something occurs. Temporal expressions may be calendar dates, times of day, or durations, such as periods of hours, days, or even centuries, and they can be precise or ambiguous. The Timex3 standard, subsumed into TimeML, provides for marking the value of the temporal expressions, which encodes the exact calendar date of an expressed time: we did not mark this information. The strict F_1 for time expressions was unusually low, at 0.40. To investigate this, we calculated both an identification F_1 and the Fleiss κ for assignment of the type of the expression (date, time, duration, or set). The identification measure was intended to capture how well annotators agreed that there was a time expression present, regardless of whethey identified the boundaries exactly the same. This measure was 0.69, which is reasonable. The κ value indicated "very good" agreement, at 0.77. These two numbers together would seem to indicate that the annotators would usually agreed that a time expression was present, and then also usually agree on the type; but that there was substantial variation regarding exact time expression boundaries. In the end, this is not a substantial problem, as time expressions were quite sparse in the data: there were only 349 time expressions identified in the whole corpus.

A time link is a relationship between two times, two events, or an event and a time. It indicates that a particular temporal relationship holds between the two, for example, they happen at the same time, or one happens for the duration of the other. Other less intuitive examples of time links between two events include if one event is temporally related to a specific subpart of another event, or imposes a truth-condition on another event. Time links fall into three major categories (Temporal, Aspectual, and Subordinating), each of which has a number of subtypes. Temporal relationships indicate a strict ordering between two times, two events, or a time and an event. Six of the temporal links are inverses of other links (e.g., *After* is the inverse of *Before*, *Includes* is the inverse of *Included By*, and so forth). Annotators used one side of the pair preferentially (e.g., *Before* was preferred over *After*), unless the type was specifically lexicalized in the text. Aspectual links indicate a relationship between an event and one its sub-parts. Subordinating links indicate relationships involving events that take arguments. Good examples are events that impose some truth-condition on their arguments, or imply that their arguments are about future or possible worlds. The strict F_1 measure for time links was 0.52, which, while seemingly poor, is actually quite good. This measure captures the *exact* syntactic match between the markings of the two annotators, not the correspondence between the final temporal partial ordering. How to measure the quality of the match between the orderings is still a matter of some debate (Tannier and Muller, 2011). In the original TimeBank annotation, exact matches were more in the 0.40-0.50 range.

3.5. Word semantics

Four different layers of word semantics were added to the corpus. First, every open-class word in the corpus was sense-disambiguated relative to Wordnet 3.0 (Fellbaum, 1998). Here, our agreement rates were surprisingly good, coming in at a strict F_1 of 0.72.

Second, multi-word expressions (MWEs) were semi-automatically annotated. Here we used the jMWE library (Finlayson and Kulkarni, 2011; Kulkarni and Finlayson, 2011) to calculate the MWEs presented to the annotators. The calculation relied on the best detector identified in those papers, which was the *Consecutive +ProperNouns +PatternInflection +MoreFrequentAsMWE* detector. That detector had a reported performance 0.83 F_1 for MWEs drawn from Wordnet, but, after correction, we obtained an F_1 of 0.65 on the N2 texts.

Third, semantic roles (Palmer et al., 2005) were semi-automatically annotated. The semantic roles were first calculated using an in-house semantic role labeler, written in Java, that was implemented based on descriptions in the literature (Gildea and Jurafsky, 2002; Pradhan et al., 2005). These annotations were then corrected by the annotators. Unfortunately, the agreement here was quite low, at a strict F_1 of 0.40. We calculated an alternative measure of agreement that only takes into account core arguments, and this F_1 measure achieves a better agreement rate of 0.59, which is reasonable. After conducting the annotation, we investigated why were obtained such low agreement rates, rereading the key papers and asking the investigators involved in those projects. The key difference, it seems, is that we

Team #	Layers
1	Word Senses, Part of Speech Tags, Lemmas, Multi-Word Expressions
2	Referring Expressions, Co-Reference Groups
3	Time Expressions, Events
4	Semantic Roles
5	Time Links
6	Referent Properties

Table 4: Six teams and the layers for which they were responsible.

were not providing our annotators with a small set of discourse segments as candidates for arguments. In the original PropBank study, these discourse segemnts were drawn directly from the CFG parses for each sentence. While we had this information available, we did not know about this strategy, and so our annotators were free to pick any argument boundaries they wanted, which led to quite a bit of variation, which suppressed the overall agreement.

Finally, properties that modified referents were marked. There were twelve property types: Physical, Material, Location, Personality, Name/Title, Origin, Ordinal, Quantification, Class, Whole, Mass Amount, and Countable Amount. There was a thirteenth property type, Descriptive, which was a catch-all if no other property type was appropriate. Annotators marked the extent of each property, its type, and associated it with the nearest relevant referring expression. The agreement on this novel representation layer was excellent: the strict F_1 was 0.87, with an identification F_1 (see above) of 0.92 and a Fleiss κ of 0.91 over the assignment of types.

4. Annotation process

The annotation was conducted by 15 annotators split across 6 teams. Each team consisted of two annotators and one adjudicator (some people worked on more than one team, or dropped out in the middle of the project), and each team was responsible for a different set of annotation layers, as shown in Table 4. Texts were split into batches of about 3000 words, and distributed to the teams on an as-needed basis, usually once every 1-3 weeks. The annotators would each annotate their assigned texts, producing two parallel sets of annotations. They would then meet with the adjudicator, sometimes in person, but more often via video conference. The adjudicator had typically worked previously an annotator on the layers in question, and was somewhat more experienced in the process of annotation and details of the layer. The adjudicator then merged the annotator texts into an adjudication text, and this text was corrected by the adjudicator in consultation with the annotators during the adjudication meeting to produce the gold standard merged text.

Annotation was carried out entirely with the Story Workbench annotation tool (Finlayson, 2008; Finlayson, 2011b). The Story Workbench is a platform for general text annotation. It is free, open-source, cross-platform, and user friendly. It provides support for annotating many different annotation layers (including all those mentioned in this paper), as well as conducting annotation in a semi-automatic fashion, where initial annotations are generated by automatic analyzers and can be corrected by human annotators. Importantly, the workbench includes a number of tools that ease the annotation process. First, the user interface incorporates a fast feedback loop for giving annotators information on annotation validity: when an annotation is syntactically invalid, or semantically suspect, a warning or error is shown to the annotator, and they are prompted to correct it. The workbench also contains a tool for automatically merging annotations from different texts into one. This tool was used not only to produce the texts that were corrected during the adjudication meetings, but also to produce the final texts included in the corpus. The workbench is extensible at many different levels, admitting new annotation layers and automatic analyzers.

Because the annotation of some layers depended on other layers being complete, annotation was organized into a two-stage process. In this process, teams 1-4 would annotate and adjudicate a text, after which teams 5 & 6 would pick up those annotated texts and annotate their layers. These texts were then merged together into the final gold standard texts that contained all layers of annotation, and whatever remaining inconsistencies were corrected by the project manager in consultation with the adjudicators. Annotation of the corpus took approximately 12 months.

5. Release of data

This paper is accompanied by an archive that contains the actual annotated files and supporting documentation. The archive may be downloaded from the MIT DSpace online library repository at the following url:

`http://hdl.handle.net/1721.1/85893`

The archive contains several different types of files. First it contains the annotation guides that were used to train the annotators. The guides are numbered to match the team numbers in Table 4. Included here are not only detailed guides for some layers, as produced by the original developers of the specification, but also our synopsis guides for each layer, which were used as a reference and further training material for the annotators. Also of interest are the general annotator and adjudicator training guides, which outline the general procedures followed by the teams when conducting annotation. Those who are organizing their own annotation projects may find this material useful.

Second, the archive contains a comprehensive manifest, in Excel spreadsheet format, listing the filenames, word counts, sources, types, and rough titles of all the texts that are part of the corpus.

Third, the archive contains copies of all the articles by the authors referred to in this paper, a penultimate version of this paper itself (lacking only the exact archive url), as well as a copy of the *Sirat Ibn Hisham*, one of the collections drawn upon for the religious texts. The raw text of all the other translations of the religious texts are available through the USC Center for Muslim-Jewish Engagement, at the following url:

`http://www.usc.edu/cmje/`

Finally, the archive contains the actual corpus data files, in Story Workbench format, an XML-encoded stand-off annotation scheme. The scheme is described in the file format specification file, also included in the archive. These files can be parsed with the aid of any normal XML reading software, or can be loaded and edited easily with the Story Workbench annotation tool, also freely available.

Unfortunately there are a number of data files that are not included in the v1.0 release of the archive. These include all 29 of the OpenSource.gov files, as well as 4 of the Inpire magazine articles. The 29 OpenSource.gov files were not included because, at the time of this writing, they are still in the process of having their *For Official Use Only* (FOUO) classification lifted. As noted previously, these files were drawn from a US government website the material of which may usually only be released to those with US military funding. We have asked for the restriction to be lifted, and are waiting for them to do so. The 4 missing Inspire articles are not included because, as the final version of this article was being submitted, we realized those files were missing the Referent Properties layer. This layer was either never annotated on those files, or the data was lost. We are in the process of having that layer re-annotated on those files.

As these files and data become available, we will continue to issue updated versions of the corpus. Links to these updated versions will be available from the original url listed above. In the case that a researcher would like access to the FOUO files before that restriction is lifted, they should get in touch with the first author (Mark Finlayson) directly.

6. Acknowledgements

The preparation of this article by Dr. Finlayson was funded by the U.S. Defense Advanced Research Project Agency (DARPA) under contract number D12AP00210. Drs. Halverson and Corman were also partially funded by DARPA under contract number D12AP00074, as well as by the Department of Defense Human Social Culture Behavior (HSCB) program under Office of Naval Research (ONR) contract number N00014-09-1-0872. We would like to thank the many scientists, engineers, and other scholars associated with DARPA's Narrative Networks (N2) program, for their input and support in collecting this data. Nevertheless, the views expressed here are solely our own, and do not necessarily reflect those of N2-affiliated researchers, DARPA, ONR, the U.S. military, or the U.S. government. We thank Chase Clow, research assistant at ASU, who helped mine stories from the OpenSource.gov data. We also thank our annotation team: project manager Jared Sprague, and annotators Julia Arnous, Wendy Austin, Valerie Best, Aerin Commins, Justin Daoust, Beryl Lipton, Josh Kearney, Matt Lord, Molly Moses, Sharon Mozgai, Zanny Perrino, Justin Smith, Jacob Stulberg, and Ashley Turner.

7. References

al Bukhari, M. (1994). *Sahih Bukhari*. Kazi Publications, Chicago. Translated by M. Muhsin Khan.

Ceran, B., Karad, R., Corman, S., and Davulcu, H. (2012). A hybrid model and memory based story classifier. In *Proceedings of the 3rd Workshop on Computational Models of Narrative*, pages 58–62, Istanbul, Turkey.

Fellbaum, C. (1998). *Wordnet: An electronic lexical database*. MIT Press, Cambridge, MA.

Finlayson, M. A. and Kulkarni, N. (2011). Detecting multi-word expressions improves word sense disambiguation. In *Proceedings of the 8th Workshop on Multiword Expressions*, pages 20–24, Portland, OR.

Finlayson, M. A., Fisseni, B., Gentner, D., Gerrig, R., Lwe, B., Joewenstein, J., Mani, I., Meister, J. C., and Young, R. M. (2013). Symposium: Computational and cognitive aspects of narratives. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, pages 81–82, Berlin, Germany.

Finlayson, M. A. (2008). Collecting semantics in the wild: The Story Workbench. In *Proceedings of the AAAI Fall Symposium on Naturally-Inspired Artificial Intelligence*, pages 46–53, Arlington, VA.

Finlayson, M. A. (2011a). Corpus annotation in service of Intelligent Narrative Technologies. In *Proceedings of the 4th Workshop on Intelligent Narrative Technologies*, pages 17–20, Stanford, CA.

Finlayson, M. A. (2011b). The Story Workbench: An extensible semi-automatic text annotation tool. In *Proceedings of the 4th Workshop on Intelligent Narrative Technologies*, pages 21–24, Stanford, CA.

Finlayson, M. A. (2014). Java libraries for accessing the Princeton Wordnet: Comparison and evaluation. In *Proceedings of the 7th Global Wordnet Conference*, pages 78–85, Tartu, Estonia.

Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, 28:245–288.

Halverson, J. R., Hoddall, H. L., J., and Corman, S. R. (2011). *Master Narratives of Islamist Extremism*. Palgrave Macmillan, London.

Hervas, R. and Finlayson, M. A. (2010). The prevalence of descriptive referring expressions in news and narrative. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 49–54, Uppsala, Sweden.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan.

Kulkarni, N. and Finlayson, M. A. (2011). jMWE: A Java toolkit for detecting multi-word expressions. In *Pro-*

ceedings of the 8th Workshop on Multiword Expressions, pages 122–124, Portland, OR.

Palmer, M., Kingsbury, P., and Gildea, D. (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105.

Pradhan, S., Hacioglu, K., Krugler, V., Ward, W., Martin, J., and Jurafsky, D. (2005). Support vector learning for semantic argument classification. *Machine Learning*, 60:11–39.

Pustejovsky, J., Castano, J., Ingria, R., Sauri, R., Gaizauskas, R., Setzer, A., and Katz, G. (2003). TimeML: Robust specification of event and temporal expressions in text. In *Proceedings of the 5th International Workshop on Computational Semantics (IWCS)*, Tilburg, The Netherlands.

Abu Dawud. (2000). *Sunan Abu Dawud*. Kitab Bhavan, New Delhi. Translated by Ahmad Hasan.

Ibn Hisham. (2000). *Sirat Ibn Hisham*. al-Falah Foundation, Cairo. Translated by Inas A. Farid.

Imam Malik. (2005). *Malik's Muwatta*. Translated by Aishah Abdarahman at-Tarhumana and Yaqub Johnson.

Muslim ibn al Hajjaj. (2000). *Sahih Muslim*. Kitab Bhavan, New Delhi. Translated by Abdul Hamid Siddiqui.

Tannier, X. and Muller, P. (2011). Evaluating temporal graphs built from texts via transitive reduction. *Journal of Artificial Intelligence Research*, 40:375–413.

Toutanova, K., Klein, D., Manning, C., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL*, pages 252–259, Edmonton, Canada.

In this paper, we describe the construction of a semantically annotated corpus of clinical texts for use in the development and evaluation of systems for automatically extracting clinically significant information from the textual component of patient records. The paper details the sampling of textual material from a collection of 20,000 cancer patient records, the development of a semantic annotation scheme, the annotation methodology, the distribution of annotations in the final corpus, and the use of the corpus for development of an adaptive information extraction system. The resulting corpus is the most richly semantically annotated resource for clinical text processing built to date, whose value has been demonstrated through its use in developing an effective information extraction system. Third, every text in the corpus has been annotated for 14 layers of syntax and semantics, including: referring expressions and co-reference; events, time expressions, and temporal relationships; semantic roles; and word senses. In cases where analyzers were not available to do high-quality automatic annotations, layers were manually double-annotated and adjudicated by trained annotators.Â Finlayson MA, Halverson JR, Corman S. The N2 corpus: A semantically annotated collection of Islamist extremist stories. In Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014. European Language Resources Association (ELRA). 2014. p. 896-902. Finlayson, Mark A. ; Halverson, Jeffry R. ; Corman, Steven. / His research focuses on representing, extracting, and using higher-order semantic patterns in natural language, especially focusing on narrative. His work intersects artificial intelligence, computational linguistics, and cognitive science.Â Finlayson, M. A., Halverson, J. R., and Corman, S. R., (2014) â€œThe N2 Corpus: A Semantically Annotated Collection of Islamist Extremist Stories,â€ in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), pp. 896â€“902. Finlayson, M. A. (2014) â€œJava Libraries for Accessing the Princeton Wordnet: Comparison and Evaluation,â€ in Proceedings of the 7th International Global WordNet Conference (GWC 2014), pp. 78â€“85.