Annotation Schemes to Encode Domain Knowledge in Medical Narratives

Wilson McCoy
Dept. of Interactive Games and Media
wgm4143@rit.edu

Cecilia Ovesdotter Alm
Dept. of English
coagla@rit.edu

Cara Calvelli
College of Health Sciences and Technology
cfcscl@rit.edu

Rui Li
Computing and Information Sciences
rxl5604@rit.edu

Jeff B. Pelz
Center for Imaging Science
pelz@cis.rit.edu

Pengcheng Shi
Computing and Information Sciences
spcast@rit.edu

Anne Haake
Computing and Information Sciences
arhics@rit.edu

Rochester Institute of Technology

Abstract

The broad goal of this study is to further the understanding of doctors' diagnostic styles and reasoning processes. We analyze and validate methods for annotating verbal diagnostic narratives collected together with eye-movement data. The long-term goal is to understand the cognitive reasoning and decision-making processes of medical experts, which could be useful for clinical information systems. The linguistic data set consists of transcribed recordings. Dermatologists were shown images of cutaneous conditions and asked to explain their observations aloud as they proceeded towards a diagnosis. We report on two linked annotation studies. In the first study, a subset of narratives were annotated by experts using a unique annotation scheme developed specifically for capturing decision-making components in the *diagnostic process* of dermatologists. We analyze annotator agreement as well as compare this annotation scheme to *semantic types* of the Unified Medical Language System as validation. In the second study, we explore the annotation of *diagnostic correctness* in the narratives at three relevant diagnostic steps, and we also explore the relationship between the two annotation schemes.

1 Introduction

From a scientific perspective, it is important to understand the cognitive decision-making processes of physicians. This knowledge can be useful for natural language processing systems and user-centered decision support in the medical field. Annotation schemes can be used to encode such information. With the growth of electronic medical records, reliable and robust annotation schemes can potentially also make the retrieval and use of archived medical information more effective. This research analyzes two annotation schemes in the context of dermatology for transcribed verbal medical narratives. One scheme is additionally compared to semantic types in the MetaMap semantic network contained in the Unified Medical Language System or *UMLS* (Aronson, 2006) as external validation. This study furthers research in linguistically annotated corpora by creating and validating schemes with future potential applications in the medical industry.

2 Data Set

For clarity, we begin by outlining the original data collection experiment (McCoy et al., 2012). The experiment included 16 physicians with dermatological expertise. Of these, 12 were attending physicians and 4 were residents (i.e., dermatologists in training). The experts were shown a series of 50 images of dermatological conditions.[1] The experts' verbal narratives were recorded, as were their eye-movements. 707 narratives were used in this study.

The participating physicians were instructed to narrate their thoughts and observations about each image to a silent student, while arriving at a differential diagnosis and possible final diagnosis. This data elicitation approach is a modified version of the Master-Apprentice interaction scenario (Beyer and Holtzblatt, 1997). The verbal data were later

[1] Some images courtesy of Logical Images, Inc.

Proceedings of the 6th Linguistic Annotation Workshop, pages 95–103,
Jeju, Republic of Korea, 12-13 July 2012.

time-aligned using the speech processing tool Praat[2] (Boersma, 2001) and stored as Praat TextGrid files. Disfluencies and pauses were also transcribed (e.g. Womack et al. (2012) analyzes certain disfluencies in this data set). The average length of a narrative is 55.6 seconds with an average of 105 words. There is an average of 15.4 pauses across narratives and an average total silent time of 19.7 seconds per narrative.

For methodological reasons, clean text transcripts were distributed to annotators in the two studies. These were cleaned of most disfluencies and agrammatical characteristics that otherwise could distract the annotator while reading.

3 Annotation Study 1: Diagnostic Thought Units

An annotation scheme was created to reveal the cognitive decision-making processes of physicians. This scheme divides the narratives into diagnostic units known henceforth as *thought units*. A thought unit is a single word or sequence of words to receive a descriptive label based on its part in the diagnostic process. With input from dermatologist and co-author Cara Calvelli, referred to below as MD 1, we defined a set of nine basic thought units. The creation of this scheme was separate from the annotation procedure. The tags and abbreviations are in Table 1.

Thought Unit Label	Tag Abbr.	Example
Patient Demographics	DEM	young
Body Location	LOC	arm
Configuration	CON	linear
Distribution	DIS	acral
Primary Morphology	PRI	papule
Secondary Morphology	SEC	scale
Differential Diagnosis	DIF	*X*, *Y* or *Z*
Final Diagnosis	DX	this is *X*
Recommendations	REC	*P* should *Q*

Table 1: Thought unit tags, their abbreviations given to experts in annotation study 1, and hypothetical examples. Thought units can span multiple words in the transcripts. For clarity, thought unit tags are in capital letters.

Of the narratives, 60 were chosen to be annotated in the first study. These represented transcripts of 10 images, selected because of their differing medical lesion morphologies. For each of the chosen images, the three longest and three shortest transcripts were included, thus comprising examples with potentially larger vs. smaller numbers of thought unit tokens (e.g. to understand which thought units were likely to be skipped).

Involving an [older patient's]$_{DEM}$ [infraorbital area]$_{LOC}$ is a [pearly papule]$_{PRI}$ with [overlying telangiectasia]$_{SEC}$ suggestive of a [basal cell carcinoma]$_{DX}$.

Figure 1: An example annotated narrative. Annotated text is shown inside of brackets followed by the annotated thought unit tag abbreviation subscript.

Printed and shuffled transcripts of the 60 narratives were independently provided to two physicians, referred to below as MD 1 and MD 2, who did not take part in the original data elicitation experiment. The expert annotators were instructed to mark sequences of words which they believed comprised the provided thought units. A short example narrative as annotated by one expert and the associated image is shown in Figure 1.

MD 2 expanded the tag set with an additional subset of thought unit tags, however, they are largely not considered in this analysis.[3] This is because of their inability to be compared to thought unit tags used by MD 1 as well as their generally low frequency (9 of the 15 new thought units each account for less than 1% of MD 2's thought unit tokens).

[2]See: `http://www.fon.hum.uva.nl/praat/`.

[3]MD 2 added the tags *Color* (COL), *Adjective* (ADJ), *Disease Category* (CAT), *Associated Skin Condition* (ASX), *Vague Skin Impression* (VSI), *Skin Morphologic Diagnosis* (SDX), *General Description* (GD), *Size* (SIZE), *Descriptive Classifier* (CLASS), *Temporal Description* (TEMP), *Underlying Diagnosis* (UDX), *Associated History* (AHX), *Underlying Medical Description* (UMD), and *Severity* (SEV).

After these annotations were completed, and after sufficient time had passed, the same set of 60 transcripts, reshuffled, were given to MD 1 again to re-annotate. MD 1 was aware that this was a re-annotation. MD 1's original annotation is referred to as MD 1a and the re-annotation as MD 1b. With the completion of this annotation set, inter-annotator and intra-annotator agreement could be analyzed.

Thought unit annotations were then time-aligned as tiers below a word tier in Praat. This allowed us to compare thought unit tokens directly along a temporal scale visually as well as automatically. It also allows the comparison of both local and global speech phenomena. Figure 2 shows a slice of a diagnostic narrative in Praat with thought unit annotations that have perfect overlap between MD 1a and MD 1b. It also shows that there was partial disagreement by MD 2 regarding the SEC token. The MD 1a and MD 1b annotations included "surrounding" as part of the secondary lesion morphology and the MD 2 annotation did not. In this example, MD 2 also partially agreed with MD 1's PRI tokens but not on the complete word sequence; "violaceous" is marked as COL, one of MD 2's added tags.

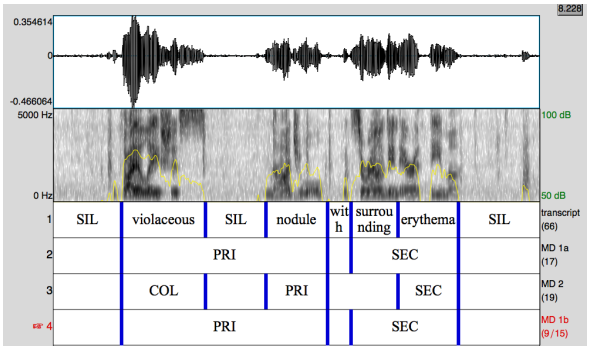

Figure 2: A screenshot of the annotation data entry.

Wordle[4] was used to visualize the prominence of concepts by thought units, given frequencies. The word clouds for *body location* (LOC) and *primary morphology* (PRI) are shown in Figure 3 and Figure 4, respectively. In Figure 3, as expected, words relating to body parts are most prominent. In Figure 4, the most prominent words, *plaque, papule, and patch*, are important primary morphology types.

[4]See `http://www.wordle.net`. In Figures 3 and 4, concepts with multiple word forms were lemmatized.

Figure 3: A word cloud generated from all words marked as *body location*.

Figure 4: A word cloud generated from all words marked as *primary morphology*.

3.1 Analysis of Thought Units' Distributions

Occurrences of each thought unit were tabulated. Raw counts as well as their percentages of the total thought unit tokens are shown in Tables 2 and 3. The percent of narratives in which a thought unit tag appeared was also calculated. A tag was considered *present* in a narrative if any annotation (MD 1a, MD 2, or MD 1b) used it at least once in said narrative.

In regards to intra-annotation variation, the MD 1a annotation used a similar number of tokens as the MD 1b re-annotation. In fact, the tags themselves are also similarly distributed, varying by at most 5% of the total tokens. In regards to inter-annotation variation, the MD 2 annotation used roughly 144% and 143% the number of tag tokens that were used by the MD 1a and MD 1b annotations, respectively. This is largely because of the additional tags that MD 2 created.

In analyzing the presence of tags, we found that every annotated narrative contained the *primary morphology* (PRI) tag type. All but two of the nine

Tag	MD 1a	% of MD 1a tags	MD 2	% of MD 2 tags	MD 1b	% of MD 1b tags	% Present
PRI	106	23	98	15	117	25	100
LOC	39	8	97	14	58	12	88
DX	42	9	71	11	32	7	86
SEC	81	17	69	10	91	19	85
DIS	51	11	9	1	29	6	66
CON	47	10	29	4	54	12	64
DIF	73	16	35	5	64	14	61
DEM	25	5	25	4	22	5	34
REC	2	<1	3	<1	2	<1	3
Total	466	100	436	65	469	100	

Table 2: Provided thought unit tags used by each annotator, the percent of all tokens with that tag, and the percent of narratives in which tags were present. 35% of MD 2's tags were self-created, see Table 3.

Tag	MD 2	% of MD 2 tags	% Present
COL	65	10	64
ADJ	62	9	64
CAT	28	4	29
ASX	26	4	36
VSI	16	2	24
SDX	6	1	10
GD	9	1	8
SIZE	6	1	8
CLASS	6	1	7
TEMP	3	<1	5
UDX	4	1	3
AHX	3	<1	3
UMD	2	<1	3
SEV	1	<1	2
Total	237	35	

Table 3: Thought unit abbreviations created by MD 2, the percent of MD 2's tokens assigned to tags, and the percent of narratives in which tags were present (see Table 2).

provided tags appeared in more than 60% of the annotated narratives. These two tags were *patient demographics* (DEM) and *recommendations* (REC).

3.2 Temporal Distribution of Thought Units in the Diagnostic Process

The positions of thought unit tokens in the narratives combining MD 1a, MD 2, and MD 1b were also calculated and are shown in Figure 5 on the next page, excluding additional thought unit tags created by MD 2. Because tokens could span several words, the time at the center of the token was used to calculate its position. This number was then normalized to a number from 0 to 1 with 0 being the beginning of the narrative and 1 being the end. Positions were rounded down to the nearest .05.

The overall temporal reasoning trajectory found seems intuitive. Doctors tend to follow a cognitive path with most DIS, DEM, CON, and LOC tokens occurring toward the beginning, followed by PRI, SEC, and DIF tokens, and concluded with DX tokens. The REC tokens appear infrequently but mostly occur at the end alongside DIF and DX tokens.

Doctors largely follow the same descriptive path of stating medical morphologies and other observable information, creating a differential diagnosis, and then choosing a final diagnosis, thus the analysis confirmed our expectations. The observed trend could also relate to traditions and training in dermatology. MD 1 and MD 2 did not know each other and received their dermatology training in different areas of the United States. We recognize that the analysis is biased towards MD 1 as that expert annotated twice.

We performed the temporal analysis on the new thought units created by MD 2, however the results were less conclusive and are therefore not included here. The created tags *Color* (COL) and *Adjective* (ADJ) largely appear near the beginning of the narrative similarly to PRI. This, and the fact that most new thought units were rare, indicate that the new thought units seemed to represent an unnecessarily

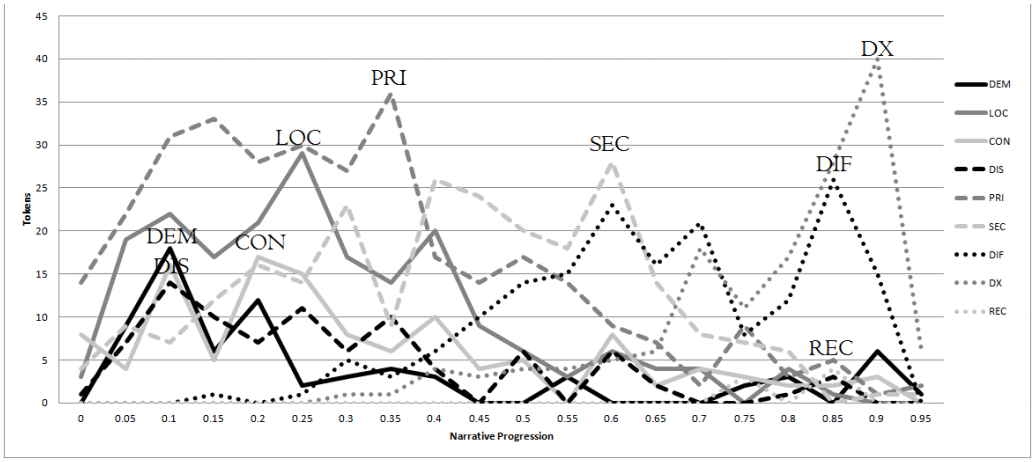

Figure 5: Distributions of provided thought unit tokens over narrative length, expressed as a ratio from 0 to 1 with 0 being the beginning and 1 being the end of the narrative. The frequency peak of each thought unit is marked.

fine granularity as a similar behavior was already captured by the provided thought units.

3.3 Agreement Metrics

Confusion matrices were created for each annotator pair. As a unit of agreement analysis, we compared overlap of tokens by individual words (including silences and disfluencies) because tokens could span and overlap in a variety of ways as shown in Figure 2. The intra-annotation matrix (MD 1a/MD 1b) is shown as a heat-map in Figure 6 with darker cells showing tags that were more often annotated together. Inter-annotation matrices were also created between MD 1a/MD 2 and MD 2/MD 1b but are not shown here. In figure 6, as a general trend, the diagonal shows that there was strong agreement on most tags. In this inter-annotation matrix, some of the most confused thought units are DIS and LOC which both refer to spatial phenomena as well as DIF and DX which both refer to diagnostic conclusions. We maintain each of these as separate labels, however, because it is good practice in dermatology to specifically assess each one.

The annotator agreement measures of observed agreement and Cohen (1960) kappa were also calculated from the data set. For the results shown in Table 4, thought units created by MD 2 were reassigned to one of the 9 provided tags based on the created confusion matrices. This was done only for this metric because MD 2 often used a created tag but in the same place as both MD 1 annotations as

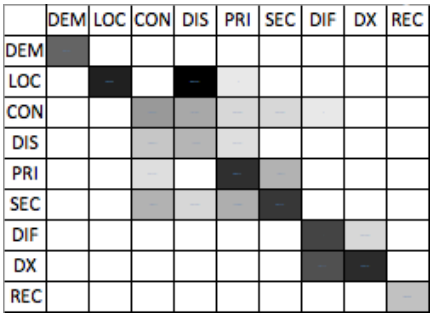

Figure 6: A heat-map of MD 1a's (columns) and MD 1b's (rows) confusion matrix. Darker cells indicate greater token overlap.

MD	1a - 2	2 - 1b	1a - 1b
% Agreement	80.69	77.72	80.98
Kappa	.56	.54	.62

Table 4: Agreement metrics for thought unit annotations. Calculations are performed pairwise for MD 1a, MD 2, and MD 1b. 1a - 1b is an intra-annotation measure.

shown in the case of COL and PRI tokens in Figure 2. With this, these metrics better represent the agreement regarding positions of tokens instead of the disagreement between the tags used. The calculations of these metrics showed moderate to good agreement between all annotation pairs.

3.4 External Validation with UMLS MetaMap

To externally validate the annotation scheme, it was compared to the semantic types used in the Uni-

fied Medical Language System (UMLS) (Bodenreider, 2004). With its 133 types, many of which are abstractions (such as "Conceptual Entity" and "Laboratory Procedure"), the UMLS ontology contains much fine-grained information. Our annotation scheme focuses on the cognitive process of dermatologists during a diagnostic procedure; we are not proposing a replacement for UMLS. Although UMLS and our annotation scheme are for different purposes (i.e., overall medicine vs. dermatology diagnostics), we regard a comparison between the two valid.

The text of each thought unit annotation was used as a query to the MetaMap semantic network. This returned a list of MetaMap entries and their semantic types. MetaMap was configured to only return the most likely match, or matches in the case of a tie. The semantic type or types of each result were counted towards the relationship to the thought unit tag the word sequence corresponded to. These relationships were then analyzed. We found that for most thought units, the most frequently occurring semantic types were often similar to the definitions of our thought units. Some examples are the LOC tag having "Spatial Concept" and "Body Part, Organ, or Organ Component" as its two most common semantic types and the DEM tag having "Age Group" and "Population Group" as its two most common semantic types.

A network density graph was created of all of these relationships with edge lengths inversely proportional to the strength of the relationship. It was too large and complex to show in this paper; instead, only the 40 strongest relationships were used to create a smaller network density graph shown in Figure 7. This also reduced noise from false positives returned by MetaMap.[5]

Based on Figure 7, a few conclusions can be drawn. PRI and SEC tags share many of the same semantic types. Eight of PRI's eleven shown relationships include semantic types that are shared among SEC's ten shown relationships. DIF has seven shown relationships compared to three of DX. Both of these thought units, however, are strongly related to "Neoplastic Process" and "Disease or Syndrome". Semantic types are also shared among DIS, LOC, and CON. These findings correspond to the confusion among these tags noted in Section 3.3 and Figure 6. Among the 40 strongest relationships, only one is not from the set of nine provided tags. This validates the tag set and indicates that perhaps *color* (COL) should be re-considered for inclusion in future work.

[5]Some noise, however, is still present. For example, the relationship between "Medical Device" and two of our created tags exists because the word 'scale' exists in a dermatological sense and as the item to weigh objects.

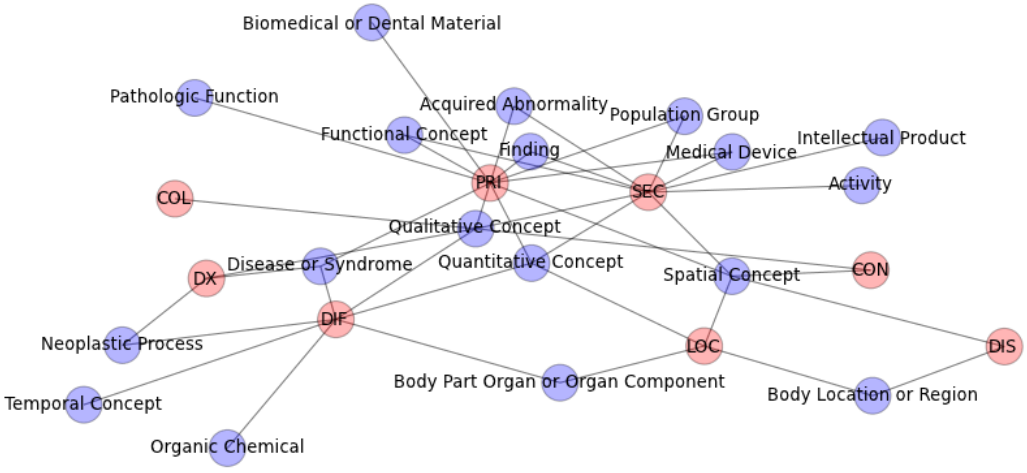

Figure 7: A network density graph of the 40 strongest relationships between text marked with thought unit tags and UMLS semantic types. The included thought units are *differential diagnosis* (DIF), *final diagnosis* (DX), *secondary morphology* (SEC), *primary morphology* (PRI), *configuration* (CON), *distribution* (DIS), *body location* (LOC), and *color* (COL) which was added by MD 2. Less strong relationships were filtered out (e.g., removing DEM, REC)

4 Annotation Study 2: Diagnostic Correctness

Cleaned transcripts were sent to three expert dermatologists referred to as MD A, MD B, and MD C to evaluate each narratives' correctness. Co-author Dr. Calvelli took part in this study as well due to limited resources and is referred to as MD A. Narratives were evaluated on three categories: correctness of the *medical lesion morphology* (*Mlm*), inclusion of the correct answer in the *differential diagnosis* (*Ddx*), and correctness of the final diagnosis (*Fdx*). Annotators were asked to use tags provided in Table 5. Inter-annotator agreements were calculated by annotator pair and are shown in Table 6. There is very good agreement between the annotators in most metrics. The lowest scores were all regarding *Mlm* most likely because of its subjectivity and greater number of class labels.

We were interested in determining how the thought units analyzed in Section 3 related to correctness annotations. To do this, we first calculated three accuracy scores for each narrative (one score for each diagnostic step scored by the annotators). The formula for correctness is shown below using Final Diagnosis (*Fdx*) as an example. Let t be a thought unit in the set T, n be a narrative in set N, and a be an annotator in set A.

$$n_{score} = \frac{\sum_{i=1}^{|A|} n(a_i(Fdx)) = \text{`Correct'} \{^{1\,:\,\text{True}}_{0\,:\,\text{False}}}{|A|}$$

We then calculated the correctness based on thought unit presence using the following formula.

Class of label	Possible labels
Medical Lesion Morphology (*Mlm*)	*Correct* *Incorrect* *None Given* *Incomplete*
Differential Diagnosis (*Ddx*)	*Yes* *No* *No Differential*
Final Diagnosis (*Fdx*)	*Correct* *Incorrect* *None Given*

Table 5: Labels for correctness annotations. To not confuse these labels with thought unit labels (Section 3), they are written with an initial capital letter and italics.[7]

Diagnostic step	Metric	A - B	B - C	C - A
Mlm	% Agr.	67.75	72.40	71.52
Ddx	% Agr.	91.84	88.46	88.71
Fdx	% Agr.	88.21	91.97	83.56
Mlm	Kappa	0.24	0.22	0.39
Ddx	Kappa	0.85	0.79	0.79
Fdx	Kappa	0.79	0.84	0.70

Table 6: Pairwise agreement metrics between MD A, MD B, and MD C performed on correctness annotations at three levels. Annotators assigned three labels to each narrative (one at each diagnostic step). See Table 5.

$$t_{score} = \frac{\sum_{i=1}^{|N|} t \text{ in } n_i \{^{n_{score}\,:\,\text{True}}_{0\,:\,\text{False}}}{\sum_{i=1}^{|N|} t \text{ in } n_i \{^{1\,:\,\text{True}}_{0\,:\,\text{False}}}$$

These scores were computed with the nine provided thought units and are shown in Table 7.

As expected, when a DX token was present, a narrative was more often marked *'Correct'* for *Fdx*. Contrary to this general finding, the appearance of a DIF token decreased the ratio of *'Correct'* tags for *Fdx*. This could be because we did not ask for a differential diagnosis in the elicitation experiment and experts generally gave differentials, so perhaps experts were more likely to give a differential if they were unsure of their diagnosis. Another interesting finding was that DEM tokens also slightly decreased the ratio of *'Correct' Fdx*. We suspect that this is because the observers were more likely to mention demographics when presented cases with which they are not as familiar.

5 Previous Work

Woods et al. (2006) performed a study to compare the UMLS vocabulary to terms used by doctors to describe images. They found that between 94% and 99% of concepts returned by the UMLS metathesaurus were regarded as exact matches by their dermatologists. The authors conclude that the UMLS metathesaurus is a reliable tool for indexing images by keywords. This provides evidence that the UMLS metathesaurus is useful as a form of validation. Hahn and Wermter (2004) have discussed the difficulties with applying natural language concepts to medical domains because of the complexity and domain-specific knowledge. Because of this we work together with expert physicians. Derma-

		Fdx		*Ddx*		*Mlm*	
Thought Unit	% Present	Present	Absent	Present	Absent	Present	Absent
PRI	100	.61	NaN	.26	NaN	.66	NaN
LOC	88	.60	.71	.29	0	.64	.81
DX	86	.66	.29	.24	.42	.67	.58
SEC	85	.67	.30	.27	.07	.70	.44
DIS	66	.69	.45	.26	.25	.63	.72
CON	64	.67	.51	.28	.16	.71	.54
DIF	61	.44	.87	.43	0	.60	.75
DEM	36	.59	.62	.38	.19	.54	.73
REC	3	.50	.61	.83	.24	.67	.66

Table 7: Ratios of correctness of the three diagnostic steps when individual thought units are present vs. when they are absent (a tag is present in a narrative if at least one annotator used it at least once in thought unit annotation). Also included are the percent of narratives in which each thought unit appeared in.

tologists were instrumental in creating schemes for annotation and several dermatologists were involved in annotating the data set. By modeling our annotation scheme after the decision-making process of a trained physician, we can better capture the domain-specific knowledge and how it is being used. Niu and Hirst (2004) have done work with annotations of clinical texts. These contain much information but do not give us insight into the cognitive process. The data set reported on in this study shows diagnostic cognitive processes through narrations spoken impromptu. Because of this, the data set captures cognitive associations, including speculative reasoning elements. Such information could be useful in a decision-support system, for instance to alert physicians to commonly confused diagnostic alternatives. Other work has been done in annotating medical texts. For example, Mowery et al. (2008) focused on finding temporal aspects of clinical texts, whereas we attempt to show the steps of the cognitive processes used by physicians during decision-making. Marciniak and Mykowiecka (2011) also report on annotating medical texts. They verified an automatic system against manual annotation of hospital discharge reports for linguistic morphologies.

Importantly, this study responds to the need identified by Kokkinakis and Gronostaj (2010) for better methods for parsing scientific and medical data. The presented annotations schemes and the annotated data set we report upon will be useful for developing and evaluating relevant systems for processing clinical dermatology texts. This research is also a starting point for empirically exploring the theoretical division of physicians' decision-making systems by Croskerry (2009) into "intuitive" and "analytical" (p. 1022). We plan to investigate the relationship between thought units and Croskerry's hypothesized differences in medical reasoning situations further.

6 Conclusion

This study investigates two annotation schemes that capture cognitive reasoning processes of dermatologists. Our work contributes to the understanding the linguistic expression of cognitive decision-making in a clinical domain and appropriate annotation processes that capture such phenomena. With this information, intuitive decision support systems and new electronic medical records storage and retrieval methods can be developed to help the growing field of medical technology. In future work, integration of gaze data will allow us to map eye-movement patterns to thought units; the multimodal approach will elucidate the link between visual perceptual and verbally expressed conceptual cognition.

Acknowledgements

This research was supported by NIH 1 R21 LM010039-01A1, NSF IIS-0941452, RIT GCCIS Seed Funding, and RIT Research Computing (http://rc.rit.edu). We thank Lowell A. Goldsmith, M.D., anonymous reviewers, Preethi Vaidyanathan, and transcribers.

References

Alan R. Aronson. 2006. MetaMap: Mapping Text to the UMLS Metathesaurus. July.

Hugh Beyer and Karen Holtzblatt. 1997. *Contextual Design: Defining Customer-Centered Systems*. Morgan Kaufmann.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, pages D267–D270.

Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, pages 341–345.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20:37–46.

Pat Croskerry. 2009. A Universal Model of Diagnostic Reasoning. *Academic Medicine*, pages 1022–1028.

Udo Hahn and Joachim Wermter. 2004. High-Performance Tagging on Medical Texts. *Proceedings of the 20th international conference on Computational Linguistics*, pages 973–979.

Dimitrios Kokkinakis and Maria Toporowska Gronostaj. 2010. Linking SweFN++ with Medical Resources, towards a MedFrameNet for Swedish. *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, pages 68–71.

Malgorzata Marciniak and Agnieszka Mykowiecka. 2011. Towards Morphologically Annotated Corpus of Hospital Discharge Reports in Polish. *Proceedings of the 2011 Workshop on Biomedical Natural Language Procesing, ACL-HLT*, pages 92–100.

Wilson McCoy, Cecilia Ovesdotter Alm, Cara Calvelli, Jeff Pelz, Pengcheng Shi, and Anne Haake. 2012. Linking Uncertainty in Physicians' Narratives to Diagnostic Correctness. *Proceedings of the ExProM 2012 Workshop*.

Danielle L. Mowery, Henk Harkema, and Wendy W. Chapman. 2008. Temporal Annotation of Clinical Text. *BioNLP 2008: Current Trends in Biomedical Natural Language Processing*, pages 106–107.

Yun Niu and Graeme Hirst. 2004. Analysis of Semantic Classes in Medical Text for Question Answering. *ACL 2004 Workshop on Question Answering in Restricted Domains*.

Kathryn Womack, Wilson McCoy, Cecilia Ovesdotter Alm, Cara Calvelli, Jeff Pelz, Pengcheng Shi, and Anne Haake. 2012. Disfluencies as Extra-Propositional Indicators of Cognitive Processing. *Proceedings of the ExProM 2012 Workshop*.

James Woods, Charles Sneiderman, Karam Hameed, Michael Ackerman, and Charlie Hatton. 2006. Using UMLS Metathesaurus Concepts to Describe Medical Images: dermatology vocabulary. *Computers in Biology and Medicine 36*, pages 89–100.