Using linguistic analysis to characterize conceptual units of thought in spoken medical narratives

Kathryn Womack\textsuperscript{1}, Cecilia Ovesdotter Alm\textsuperscript{2}, Cara Calvelli\textsuperscript{3}
Jeff B. Pelz\textsuperscript{4}, Pengcheng Shi\textsuperscript{5}, Anne Haake\textsuperscript{5}

\textsuperscript{1} Dept. of ASL & Interpreting Edu., \textsuperscript{2} Dept. of English, \textsuperscript{3} College of Health Sciences & Tech.
\textsuperscript{4} Center for Imaging Science, \textsuperscript{5} Computing & Information Sciences
Rochester Institute of Technology, Rochester, NY, USA
kaw8159@rit.edu, coagla@rit.edu, cfcscl@rit.edu
pelz@cis.rit.edu, spcast@rit.edu, arhics@rit.edu

Abstract

This study explores spoken medical narratives in which dermatologists were shown images of dermatological conditions and asked to explain their reasoning process while working toward a diagnosis. This corpus has been annotated by a domain expert for information-rich conceptual units of thought, providing opportunity for analysis of the link between diagnostic reasoning steps and speech features. We explore these annotations in regards to speech disfluencies, prosody, and type-token ratios, with the finding that speech tagged within thought units is unique from non-tagged speech in each of these aspects. Additionally, we discuss pattern differences in temporal thought unit distribution based on diagnostic correctness.

Index Terms: conceptual units of thought, speech features, disfluencies, diagnostic correctness, medical reasoning

1. Introduction

In expert domains such as the medical field, high-stakes cognitive reasoning and decision-making take place. This study aims to further our understanding of dermatological diagnostic decision making as expressed in speech. We use a spoken corpus with manual annotations by an expert for diagnostic units of thought (henceforth, thought units), i.e. semantic units which capture specific conceptual steps of the diagnostic reasoning process, such as describing the form or distribution of a lesion.

We look at the patterns of speech disfluencies, prosody, and type-token ratios with respect to diagnostic thought units in the medical narratives. We also describe findings of differences in patterns of temporal trajectory of thought unit tags based on whether the given final diagnosis is correct or incorrect.

Analysis of linguistic phenomena, such as speech disfluencies, in this specialized medical domain is useful to better understand how they vary in different situations and if they can be a meaningful linguistic window toward better understanding the medical reasoning process. For example, studies have shown that disfluencies indicate cognitive processing (e.g. \cite{1, 2, 3}), but such research thus far has largely focused on casual or conversational discourse. We investigate if there is linguistic evidence suggesting that the information coded by thought units can be distinguished from other speech in the narratives, since they reflect critical steps in the medical decision-making process that may require greater cognitive demand. Identifying these patterns is also a step towards automatic labeling of diagnostic thought units or other similar semantic phenomena \cite{4} and spoken language understanding \cite{5, 6}.

We find evidence that patterns of speech within these information-rich thought units differ in character from the surrounding speech and are unique compared with previous work with casual speech corpora. This finding is a step toward being able to identify, with the help of linguistic evidence from speech, where reasoning errors such as misdiagnoses occur.

2. Previous work

Researchers have considered the relevance of patterns of disfluencies for examining cognitive load from a linguistic perspective, but more work is needed that takes advantage of such speech features, and other patterns of speech, as a window to understanding the medical reasoning process. For example, filled pauses tend to precede topics that are being introduced to the discourse or are relatively complex \cite{1, 7, 8, 9}. Disfluencies have also been found useful in information extraction and segmentation of speech \cite{4, 10}.

In related work, silent pauses are not typically considered as a speech disfluency. Some studies have looked at silent pauses’ distribution and patterns \cite{11, 12} and the perception of pauses by listeners \cite{13, 14}, but silent pauses are not typically treated as being in the same category as filled pauses or repairs.

However, it has been suggested that listeners seem to better comprehend words or phrases following disfluencies because of the pause before the word, not necessarily because of the disfluency itself; and that this effect holds equally for filled or silent pauses of the same duration \cite{15, 16}. Comparison of event-related potential (ERP) work by MacGregor et al. \cite{17} suggests that cognitive effects of disfluent silent pauses on listeners are the same as filled pauses \cite{18}, but that the effects of disfluent repetitions differ \cite{19}. While it is unlikely that silent pauses function identically to other disfluencies, their production and reception need to be further examined and better understood. For this reason, this study treats silent pauses as a type of disfluency.

Improving our understanding of the relationship between linguistic features and medical reasoning has positive implications for understanding medical reasoning in a new way and for helping to identify, with speech features, where this process breaks down and leads to an incorrect diagnosis. Speech features in this corpus have been able to predict diagnostic correctness to an initial encouraging degree of accuracy, as reported by McCoy et al. \cite{20}, but further analysis is necessary to improve our understanding of the underlying processes.
The current model of medical decision making describes two types of reasoning: the "intuitive" process which relies on pattern matching based on the physicians' experience and the step-by-step "analytic" process which carefully considers all information [21]. These systems can break down, or the system used may be inadequate for a number of reasons. This study's findings can lead toward predicting which reasoning process a physicians are using and where in the reasoning process they might make a wrong turn that leads to an incorrect diagnosis.

3. Data and annotation

The corpus used in this paper consists of transcripts acquired from a data elicitation scenario involving 16 dermatologists. Each dermatologist was shown 50 images of dermatological conditions and asked to describe the image while working toward a diagnosis. This study uses a subset of the narratives which have been specially annotated and are discussed below.

The data elicitation used a modification of the Master-Apprentice scenario described by Beyer and Holtzblatt [22]. A student "apprentice" sat silently with an expert dermatologist "master" while the "master" narrated his or her thoughts. The narratives are monologues; however, this scenario creates the feeling of dialogue in the experts and encourages narration of their thoughts in rich detail.

This study considers the narratives' speech, transcripts, and annotations. The narratives were manually single-annotated and time-aligned at the word level using Praat [23]; Figure 1 shows an example. Four transcribers worked on time-alignment with supervision. Every speech token was transcribed, including disfluencies and silent pauses over 0.3 seconds. For methodological reasons, analysis considers silent pauses that occur after the physician started speaking and before they finished speaking.

The thought unit annotation scheme encodes narratives using nine thought units, listed in Table 1, which attempt to identify common, important steps in the diagnostic process; this annotation scheme is discussed in detail by McCoy et al [24]. Initially, co-author Calvelli and a second dermatologist annotated 60 narratives. This allowed evaluation of inter- and intra-annotator agreement, which was moderate to good [24]. For this study, the annotated subcorpus was expanded to include an additional 157 narratives annotated by Dr. Calvelli, leading to a total of 217 annotated narratives. Current analysis accordingly considers Dr. Calvelli's initial annotation and the extended annotated subcorpus (N=217).

There were two differences between the initial and the more recent annotation scenarios. The first is that the transcripts given to the annotators for the original 60 narratives had been mostly cleaned of disfluencies and given punctuation for readability, while the second set of transcripts included all tokens and had no punctuation; cleaning was not needed since no project-external annotator was included. The second change was a slight redefining of the primary and secondary medical lesion morphologies. Thus, for correspondence, this work considers these two tags as one merged unit: Morphology (MOR).

The considered subset of narratives with thought units totals almost 3 hours of audio or about 18,200 tokens; the narratives are, on average, 45 seconds. Slightly more than one-third of tokens are tagged in thought units. Individual tags, on average, span 2.6 tokens and 1.3 seconds. Labels vary in frequency, with MOR tags being the most frequent (~1,000 tags over ~2,000 tokens), then DIF (~400 tags over ~1,300 tokens), and REC being the least frequent (11 tags over 62 tokens).

Additional annotation for correctness was done on the full corpus by three annotators [24]. In this study, we consider a narrative to have a correct diagnosis if 2 or 3 of the annotators agree that the final diagnosis is correct.

4. Results and discussion

4.1. Disfluencies and thought units

Several common types of speech disfluency are examined here (with examples in Table 2): silent pauses ("SIL.") (1), non-nasal filled pauses ("ah, er, and uh") (2), nasal filled pauses ("hm and um") (3), repetitions (4), and edits (5). For this study, a repetition is a one- or two-word phrase that is repeated with a silent pause, filled pause, or nothing between the words; e.g. they uh they show. An edit is a word that was cut off; e.g. on the r- left wrist.

Table 1: List of considered thought unit labels. Primary and Secondary refer to medical lesion morphology. [25]

<table>
<thead>
<tr>
<th>Thought unit</th>
<th>Abbr.</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>DEM</td>
<td>elderly</td>
</tr>
<tr>
<td>Configuration</td>
<td>CON</td>
<td>annular</td>
</tr>
<tr>
<td>Distribution</td>
<td>DIS</td>
<td>bilaterally</td>
</tr>
<tr>
<td>Location</td>
<td>LOC</td>
<td>elbow</td>
</tr>
<tr>
<td>Primary</td>
<td>PRI</td>
<td>plaque</td>
</tr>
<tr>
<td>Secondary</td>
<td>SEC</td>
<td>crusted</td>
</tr>
<tr>
<td>Differential Diagnosis</td>
<td>DIF</td>
<td>this could be a or b</td>
</tr>
<tr>
<td>Final Diagnosis</td>
<td>DX</td>
<td>this is a</td>
</tr>
<tr>
<td>Recommendation</td>
<td>REC</td>
<td>needs a biopsy</td>
</tr>
</tbody>
</table>

Table 2: Example of disfluency types in the corpus. (Taken from a similar elicitation situation and used with permission.)

A list of the disfluency types and proportions are shown in Figure 2. Silent pauses comprised the majority of disfluencies, followed by filled pauses, with repetitions and edits having the smallest proportions.

A total of 25% of tokens in the subcorpus were disfluent; 17% of all tokens were silent pauses and about 6.5% filled pauses. This percentage of disfluencies is higher than that reported by several corpora of casual conversations (e.g. in a range of 3.5 to 6 disfluencies/100 words in casual conversations.

Figure 1: Screenshot of Praat, used to time-align each narrative's speech with its transcription and annotations, and to extract speech information for analysis.
meaning that the description of morphology is the first cognitively demanding clinical reasoning step they will take. Importantly, in the full corpus of 800 narratives, 62% of narratives in which the physician provided a correct medical morphology had a correct final diagnosis, while only 37% of narratives with an incorrect or missing morphology had a correct final diagnosis. Correctly identifying the morphology is a vital step that guides the clinicians to hone in on specific potential diagnoses [25].

Thus, analysis indicates that tagged and non-tagged text behave differently. This could be in part because thought unit-tagged text is often short noun phrases. Additionally, disfluencies are likely uttered earlier in the narrative as cognitive clues related to a concept mentioned later on. Disfluencies tend to precede cognitively demanding steps (cf. section 2), so one would expect that more disfluencies would be found preceding tagged speech. Gaze analysis, of eye-tracking data co-collected with speech in this corpus, has found that people tend to look at lesion features before they discuss them [30]. Similarly, speech phenomena that occur earlier in the narrative may be indicative of cognitive reasoning that is related to utterances later on.

### 4.2. Type versus token ratio and thought units

The type/token ratio (TTR) can indicate a text’s stylistic properties. In this data, the average TTR of the narratives is 0.6, or 0.8 if disfluencies are removed; the average TTR of non-tagged text only is very similar. In comparison, the average TTR over each type of thought unit tag is 0.9, and did not change when disfluencies were removed; as discussed above, disfluencies are less prevalent in thought unit spans. The high TTR of thought unit tags indicates they are rich and dense in vocabulary, likely because the vocabulary is specialized and the typical short span of thought unit tags. There was little variation in the average TTR for each specific thought unit tag, ranging from 0.9 to ~1. These findings indicate that thought units and non-tagged text are qualitatively different, supporting the hypothesis that thought unit tags capture specific information and behave differently than non-tagged text.

### 4.3. Prosody and thought units

<table>
<thead>
<tr>
<th></th>
<th>Tagged</th>
<th>Non-tagged</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration (mean)</td>
<td>0.48</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>(st. dev.)</td>
<td>0.31</td>
</tr>
<tr>
<td>Pitch (mean)</td>
<td>66 Hz</td>
<td>63 Hz</td>
</tr>
<tr>
<td></td>
<td>(st. dev.)</td>
<td>28</td>
</tr>
<tr>
<td>Intensity (mean)</td>
<td>148 dB</td>
<td>135 dB</td>
</tr>
<tr>
<td></td>
<td>(st. dev.)</td>
<td>55</td>
</tr>
</tbody>
</table>

Table 3: Prosody features in tokens tagged with thought units and non-tagged tokens. Each test was significant, p<0.01.

Additional analysis tested the duration, mean pitch, and mean intensity of individual tokens tagged with thought units against non-tagged tokens. As shown in Table 3, Tokens' 2TTR measures the number of unique types divided by the number of total tokens. For each narrative, all words within each thought unit tag over the whole narrative were considered for TTR measurement, even if the tags were split; e.g. “[elderly male]DEM ...

Footnotes:
1. The span of non-tagged text was not expanded, to avoid overlap. If two tags were adjacent, tokens were assigned only to their actual tag. If one token was between two tags (e.g. "SIL" in Figure 1), the token was assigned to the left tag.

2. TTR measures the number of unique types divided by the number of total tokens. For each narrative, all words within each thought unit tag over the whole narrative were considered for TTR measurement, even if the tags were split; e.g. “[elderly male]DEM ...

3. Features were extracted with Praat and tested with two-sample t-tests assuming unequal variances.
mean pitch and mean intensity were both higher for thought unit-tagged tokens, which could be an indication of prominence or importance in tagged speech [31], although further analysis would be needed. The typical short span of thought unit tags and specific parts of speech captured by them also plays a role in these results. Variance in the data was higher for every test in non-tagged tokens, showing additional evidence that thought unit-tagged tokens have less variability than non-tagged tokens.

5. Diagnostic correctness and thought units

We visually examined the number of each thought unit tag along the narratives’ temporal progress, comparing narratives in which the observer gave the correct final diagnosis against narratives with an incorrect final diagnosis, shown in Figure 4. There are several differences in the temporal trajectories of thought units that could provide clues as to the thought processes of the physicians in these contexts.

In correct narratives, the number of MOR tags peak between 25% and 60% of the way through the narrative. In contrast, MOR tags in incorrect narratives peak around 20% and then trail off. In incorrect narratives, it could be that the physicians paid less attention to the Morphology description, giving the descriptions earlier and being less likely to revisit them, than in the correct narratives, where they might have been more likely to provide evidence for their favored diagnosis.

The temporal distribution of DX vs. DIF tags also differs. In correct narratives, few DIF tags are present and many DX tags are given at the end. In incorrect narratives, there are many DIF tags during the last quarter of the narrative and few DX tags. If participants were more confident when they were correct, they may have chosen to either omit or give a smaller differential and provided a final diagnosis more readily. When incorrect, participants might have been unsure and given a longer differential instead of settling on one diagnosis.

No real difference was found in the duration of each thought unit in correct versus incorrect narratives, with the exception of MOR tags. Morphology tags in correct narratives on average lasted for 9% of the narrative’s duration, while they lasted for only 5% of incorrect narratives’ duration ($p<0.01$). This again indicates the importance of identifying the medical morphology as a critical medical reasoning step, although it is difficult to determine the relationship between morphology and diagnostic correctness. It could be that physicians who spend more time and provide more detail on morphology notice more, giving them additional clues for the final diagnosis; or it could be that physicians who give the correct diagnosis are more familiar with that diagnosis’ morphology and can explain more or draw on this experience.

6. Conclusions

This study examines information-rich thought unit annotation in spoken medical narratives, with the finding that speech tokens tagged in thought units linguistically behave distinctly from non-tagged tokens in various ways, and that the temporal patterns of thought units are different in narratives with a correct diagnosis vs. an incorrect diagnosis. These findings are a step toward characterizing information-rich speech in clinical decision-making, information extraction of semantic thought units, and identification of misdiagnoses.

In particular, several intriguing results have surfaced relating to the medical lesion morphology step, which describes the form of the lesion, in the diagnostic process. Morphology tags have the highest percentage of silent pauses and an above-average percentage of total disfluencies. The duration of Morphology tags in correct narratives is significantly longer than in incorrect narratives and they appear to show a difference in temporal trajectory between correct and incorrect narratives. Moreover, based on the full corpus, we know that physicians who give the correct medical morphology are more likely to provide the correct diagnosis. These findings point toward the need for further analysis of this critical medical reasoning step.

Further work will continue to explore the relationship between medical reasoning and linguistic patterns in medical narratives based on diagnostic correctness, physician expertise, and image characteristics, such as perceived vs. actual difficulty of a diagnostic case or type of lesion. We have also, more recently, collected a similar, larger corpus in which physicians were additionally asked to provide their level of confidence in their final diagnosis, which provides the opportunity to explore another facet of diagnostic reasoning.

7. Acknowledgements

This work was partially supported by NIH grant 1 R21 LM010039-01A1 and NSF grant IIS-0941452. We appreciate the contributions of the participants, annotators, reviewers, and the research group; and Logical Images, Inc. for images.

---

4Instructions to the participants during data elicitation were flexible to allow more natural descriptions, and physicians were not specifically asked to provide a differential diagnosis.

5A two-sample t-test assuming unequal variances was used.
8. References


