Towards Automatic Annotation of Clinical Decision-Making Style

Limor Hochberg[1] **Cecilia O. Alm**[1] **Esa M. Rantanen**[1]
Qi Yu[2] **Caroline M. DeLong**[1] **Anne Haake**[2]
1 College of Liberal Arts 2 College of Computing & Information Sciences
Rochester Institute of Technology
lxh6513|coagla|emrgsh|qi.yu|cmdgsh|anne.haake@rit.edu

Abstract

Clinical decision-making has high-stakes outcomes for both physicians and patients, yet little research has attempted to model and automatically annotate such decision-making. The dual process model (Evans, 2008) posits two types of decision-making, which may be ordered on a continuum from *intuitive* to *analytical* (Hammond, 1981). Training clinicians to recognize decision-making style and select the most appropriate mode of reasoning for a particular context may help reduce diagnostic error (Norman, 2009). This study makes preliminary steps towards detection of decision style, based on an annotated dataset of image-based clinical reasoning in which speech data were collected from physicians as they inspected images of dermatological cases and moved towards diagnosis (Hochberg et al., 2014). A classifier was developed based on lexical, speech, disfluency, physician demographic, cognitive, and diagnostic difficulty features. Using random forests for binary classification of intuitive vs. analytical decision style in physicians' diagnostic descriptions, the model improved on the baseline by over 30%. The introduced computational model provides construct validity for decision styles, as well as insights into the linguistic expression of decision-making. Eventually, such modeling may be incorporated into instructional systems that teach clinicians to become more effective decision makers.

1 Introduction

Diagnostic accuracy is critical for both physicians and patients, but there is insufficient training on clinical decision-making strategy in medical schools, towards avoiding diagnostic error (Graber et al., 2012; Croskerry & Norman, 2008). Berner and Graber (2008) estimate that diagnostic error in medicine occurs at a rate of 5-15%, and that two-thirds of diagnostic errors involve cognitive root causes.

The dual process model distinguishes between *intuitive* and *analytic* modes of reasoning (Kahneman & Frederick, 2002; Evans, 1989). Use of the intuitive system, while efficient, may lead to cognitive errors based on heuristics and biases (Graber, 2009). Croskerry (2003) distinguished over 30 such biases and heuristics that underlie diagnostic error, including anchoring, base-rate neglect, and hindsight bias.

Hammond's (1981) *Cognitive Continuum Theory* proposes that decision-making lies on a continuum from intuitive to analytical reasoning. Intuitive reasoning is described as rapid, unconscious, moderately accurate, and employing simultaneous use of cues and pattern recognition (Hammond, 1981). Analytical decision-making is described as slow, conscious, task-specific, more accurate, making sequential use of cues, and applying logical rules (Hammond, 1996). Much reasoning is *quasirational*: between the two poles of purely intuitive and purely analytical decision-making (Hamm, 1988; Hammond, 1981).

Cader et al. (2005) suggested that cognitive continuum theory is appropriate for the evaluation of decision-making in medical contexts. The current study links to another work (Hochberg et al., 2014), where the cognitive continuum was applied to physician decision-making in dermatology. Decision style was manually assessed in physician verbalizations during medical image inspection. Figure 1 shows the 4-point annotation scheme, ranging from intuitive to analytical; the two intermediate points on the scale reflect the presence of both styles, with intuitive (*BI*) or analytical (*BA*) reasoning more prevalent.

LAW VIII - The 8th Linguistic Annotation Workshop, pages 129–138,
Dublin, Ireland, August 23-24 2014.

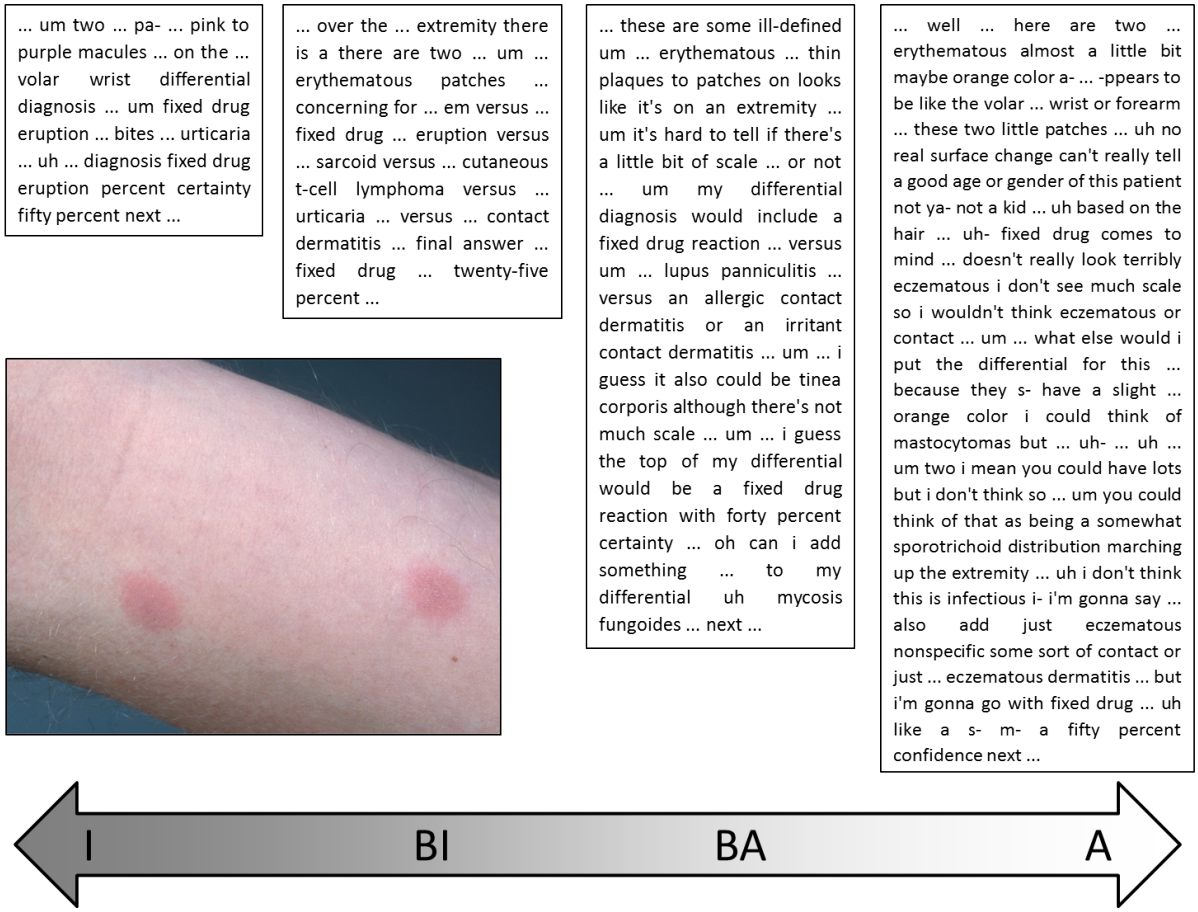

Figure 1: Four narratives along the intuitive-analytical decision-making continuum, for which annotators agreed on their labels, where *I=Intuitive, BI=Both-Intuitive, BA=Both-Analytical, A=Analytical.* The narratives were produced by different physicians for the same image case (left, used with permission from Logical Images, Inc.), and all four physicians were correct in their final diagnosis. (Confidence mentions were removed in narratives presented to annotators, to avoid any potential bias.)

This work describes computational modeling for automatic annotation of decision style using this annotated dataset, on the basis of linguistic, speaker, and image case features.

1.1 Contributions

To date, this appears to be the first study attempting to computationally predict physician decision style. Similar to the case of affect, automatic annotation of decision style can be characterized as a subjective natural language processing problem (Alm, 2011). This adds special challenges to the modeling process. Accordingly, this work details a thorough process for moving from manual to automatic annotation.

This study contributes to cognitive psychology, annotation methodology, and clinical computational linguistic analysis. Methodologically, the study details a careful process for selecting and labeling manually annotated data for modeling in the realm of subjective natural language phenomena, thus addressing the need for their characterization (Alm, 2011). Theoretically, acceptable annotator reliability on decision style, along with successful computational modeling, will lend construct validity to the dual process model. From a linguistic perspective, the identification of discriminative features for intuitive and analytical reasoning provides a springboard for further studying decision-making using language as a cognitive sensor.

Practically, prediction of decision style would also be useful for determining whether individuals are using the appropriate style for a particular task, based on analyses linking decision style to task performance. Importantly, detection of decision style from observable linguistic behaviors allows for objective measurement that avoids biases present in self-report surveys (Sjöberg, 2003; Allinson & Hayes, 1996).

2 Data and Manual Decision Style Annotation

The annotated corpus used in this study was introduced in Hochberg et al. (2014), which also discusses the manual annotation scheme and annotator strategies in greater detail. For clarity, the dataset and annotation scheme are described here briefly.

The dataset consisted of spoken narratives collected from 29 physicians as they examined 30 clinical images of dermatological cases, for a total of 867[1] narratives. Physicians described their reasoning process as they advanced towards a diagnosis, and they also estimated their confidence[2] in their final diagnosis. Narratives were assessed for correctness (based on final diagnoses) and image cases were evaluated for difficulty by a practicing dermatologist.[3]

For the manual annotation of decision style, anonymized text transcripts of the narratives were presented to two annotators with graduate training in cognitive psychology.[4] Analytical reasoning considers more alternatives in greater detail. Thus, it was expected to be associated with longer narratives, as Figure 1 illustrates. Therefore, annotators were asked not to use length as a proxy for decision style.

Narratives were randomized to ensure high-quality annotation, and 10% of narratives were duplicated to measure intra-annotator reliability. For analysis, primary ratings were used, and secondary ratings (on duplicated narratives) were used to measure intra-annotator consistency. The kappa scores and proportion agreement, detailed below, motivate the labeling and data selection process used for classification and modeling in this work.

Figure 2 shows the distribution of annotation labels for both annotators, respectively, for the whole dataset, on the original 4-point scale. In comparison, Figure 3 shows the annotators' distributions across a collapsed 2-point scale of intuitive vs. analytical, where, for each annotator, narratives labeled *BI* were assigned to *I* and those labeled *BA* assigned to *A*.

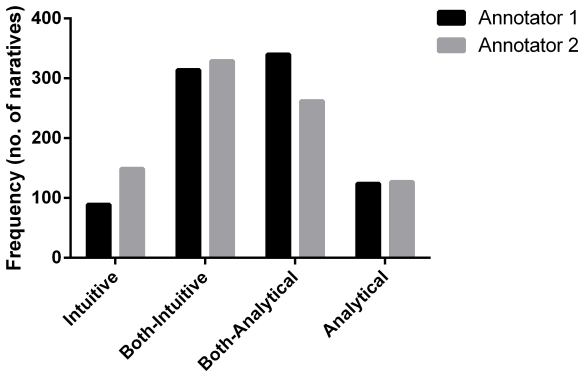

Figure 2: The distribution of ratings among the decision-making spectrum, on a 4-point scale.

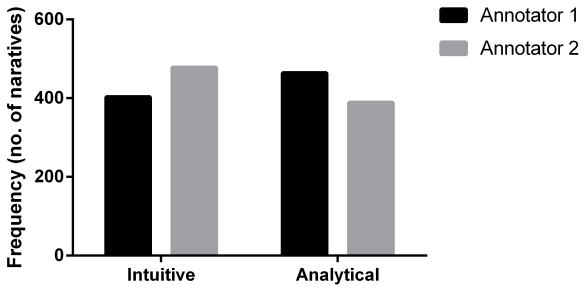

Figure 3: The distribution of ratings among the decision-making spectrum, on a 2-point scale.

Annotator agreement was well above chance for both the 4-point (Figure 4) and 2-point (Figure 5) scales. Notably, the annotators were in full agreement or agreed within one rating for over 90% of narratives on the original 4-point scale. This pattern of variation reveals both the fuzziness of the categories and also that the subjective perception of decision-making style is systematic.

Annotator agreement was also assessed via linear weighted kappa scores (Cohen, 1968). As shown in Figure 6, inter-annotator reliability was moderate, and intra-annotator reliability was moderate (Annotator 2) to good (Annotator 1); see Landis and Koch (1977) and Altman (1991).

Since both proportion agreement and kappa scores were slightly higher for the 2-point scale, the automatic annotation modeling discussed below used this binary scale. In addition, the distribution of

[1] One narrative was excluded due to extreme brevity, and two physicians each skipped an image during data collection.

[2] For consistency, this paper uses the term *confidence*, treated as interchangeable with *certainty* and similar synonymous expressions used by clinicians in the medical narratives, such as *sure*, *certain*, *confident*, just certainty percentages, etc.

[3] Some imperfections may occur in the data, e.g., in transcriptions, difficulty ratings, or annotations (or in extracted features).

[4] Annotator instructions included decision style definitions, a description of the 4-point scale and example narratives. Annotators were asked to focus on decision style as present in the text rather than speculate beyond it.

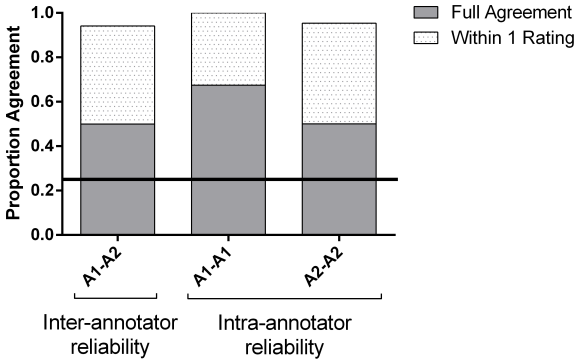

Figure 4: Inter- and intra-annotator reliability for the 4-point scheme, by proportion agreement. The reference line shows chance agreement (25%). *(A1=Annotator 1; A2=Annotator 2).*

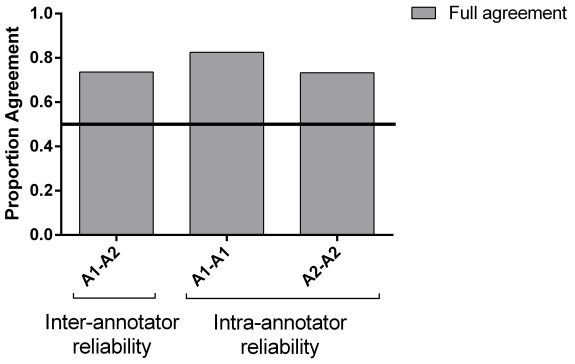

Figure 5: Inter- and intra-annotator reliability for the 2-point scheme, by proportion agreement. The reference line shows chance agreement (50%). *(A1=Annotator 1; A2=Annotator 2).*

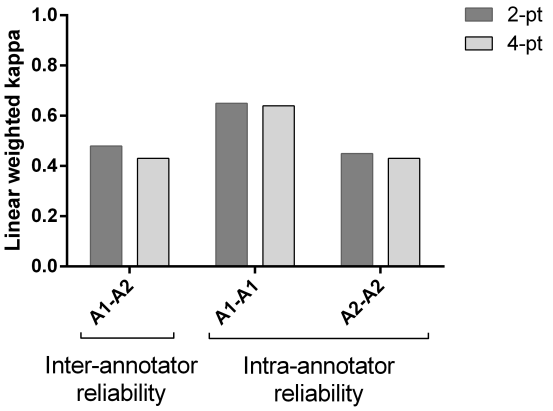

Figure 6: Annotator reliability, as measured by linear weighted kappa scores on the 2-pt and 4-pt scales.

data across binary classes was more balanced compared to the 4-point scale, as shown by the contrast between Figures 2 and 3, further making it a suitable starting point for computational modeling.

2.1 Data Selection and Labeling for Computational Modeling

This section details the systematic method used to select data for model development. The goal of the work was to develop a computational model that could automatically annotate narratives as intuitive or analytical, based on lexical, speech, disfluency, physician demographic, cognitive, and diagnostic difficulty features. The study employed a supervised learning approach, and since no real ground truth was available, it relied on manual annotation of each narrative for decision style. However, annotators did not always agree on the labels, as discussed above. Thus, strategies were developed to label narratives, including in the case of disagreement (Figure 7).

The dataset used for modeling consisted of 672 narratives.[5] Annotators were in full agreement for 614 ratings on the binary scale of intuitive vs. analytical (Figure 8).[6] Next, 49 narratives were assigned a binary label based on the center of gravity of both annotators' primary ratings (Figure 9). For example, if a narrative was rated as *Intuitive* and *Both-Analytical* by Annotators 1 and 2, respectively, the center of gravity was at *Both-Intuitive*, resulting in an *Intuitive* label. Finally, 9 narratives were labeled using the annotators' secondary ratings,[7] available for 10% of narratives, to resolve annotator disagreement.[8]

[5]Within a reasonable time frame, the text data are expected to be made publicly available.

[6]Excluding also narratives lacking confidence or correctness information.

[7]Collected to measure intra-annotator reliability.

[8]For example, if the primary ratings of Annotator 1 and Annotator 2 were *Both-Analytical* and *Both-Intuitive*, respectively, but both annotators' secondary ratings were intuitive (e.g., *Both-Intuitive* or *Intuitive*), the narrative was labeled *Intuitive*.

Narratives with disagreements that could not be resolved in these ways were excluded. As perception of decision-making style is subject to variation in human judgment, this work focused on an initial modeling of data which represent the clearer-cut cases of decision style (rather than the disagreement gray zone on this gradient perception continuum). From the perspective of dealing with a subjective problem, this approach enables an approximation of ground truth, as a validation concept.[9]

Figure 7: Narrative labeling pipeline. 614 narratives were labeled due to full binary agreement, and center-of-gravity and secondary rating strategies were used to label an additional 58 narratives for which annotators were not in agreement.

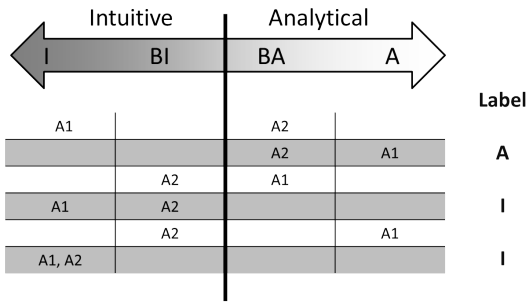

Figure 8: Demonstration of initial corpus labeling, in which 614 narratives were labeled on the basis of binary agreement.

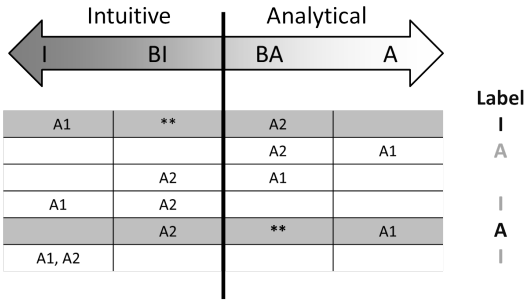

Figure 9: Demonstration of center-of-gravity strategy, used to label an additional 49 narratives.

2.2 Relationship Between Physicians' Diagnostic Correctness and Decision Style

Using the 672 narratives selected for modeling, Table 1 shows the relationship of physicians' diagnostic correctness by decision style (intuitive vs. analytical on a binary scale).

	Correct	**Incorrect**	**Total**
Intuitive	158	186	344
Analytical	106	222	328
Total	264	408	672

Table 1: Distribution of diagnostic correctness by decision style.

Overall, there was a slightly higher prevalence of intuitive reasoning, and there were more incorrect than correct diagnoses.[10] Table 1 also suggests a relationship between correctness and decision-making style, where for correct diagnoses, intuitive reasoning was more dominant. The opposite trend held for incorrect diagnoses: analytical reasoning was more frequent. Indeed, a chi-square test revealed a significant relationship between correctness and decision style, $\chi^2(1, N = 672) = 13.05, p < 0.01$.

This pattern is in line with claims that intuitive reasoning is linked to better performance when much information is to be processed; mechanisms of intuitive reasoning and pattern recognition allow individuals to overcome the limitations of their working memory (Evans, 2008). However, others have linked intuitive reasoning to decreased diagnostic accuracy, as intuitive reasoning may be prey to inappropriate

[9]Modeling of fuzzier, hard to label data, is left to future work. One possible approach is to learn the labels by using a k-nearest neighbor classifier, which identifies the most similar narratives and uses their labels to make the prediction.

[10]Contributing factors to the proportion of incorrect diagnoses might include case difficulty levels in the experimental scenario, and that physicians did not have access to additional information, such as patient history or follow-up tests.

heuristics and biases (Croskerry, 2003). Viewed from the perspective of cognitive continuum theory, the higher prevalence of incorrect diagnoses may be due to the use of decision styles that were not suited to the task demands of the particular case (Hammond, 1981). Finally, it might be the case that diagnostic difficulty was a moderating variable, where physicians preferred intuitive reasoning for less challenging cases, and analytical reasoning for more difficult cases.

3 Methods

A model was developed for the binary prediction case (intuitive vs. analytical), since the 2-point rating scheme had slightly higher annotator agreement (see Section 2). Model development and analysis were performed using the WEKA data mining software package (Hall et al., 2009). The dataset was split into 80% development and 20% final test sets (Table 2).[11] Parameter tuning was performed using 10-fold cross-validation on the best features in the development set.[12]

	80% Development Set	**20% Final Test Set**
Intuitive	276 (51%)	68 (51%)
Analytical	263 (49%)	65 (49%)
Total	539	133

Table 2: Class label statistics.

3.1 Features

Three feature types were derived from the spoken narratives to study the linguistic link to decision-making style: lexical (37), speech (13), and disfluency (3) features. Three other feature types relevant to decision-making were demographic (2), cognitive (2), and difficulty (2) features (Table 3).

Type	**Feature**	**Description / *Examples***
Lexical	exclusion	*but, without*
	inclusion	*both, with*
	insight	*think, know*
	tentative	*maybe, perhaps*
	cause	*because, therefore*
	cognitive process	*know, whether*
	. . .	
Speech	speech length	number of tokens
	pitch	min, max, mean, st. dev., time of min/max
	intensity	min, max, mean, st. dev., time of min/max
Disfluency	silent pauses	number of
	fillers	*like, blah*
	nonfluencies	*uh, um*
Demographic	gender	male, female
	status	resident, attending
Cognitive	confidence	percentage
	correctness	binary
Difficulty	expert rating	ordinal ranking
	% correctness/image	percentage

Table 3: Six feature types. The listed lexical features are a sub-sample of the total set.

Relevant *lexical* features were extracted with the Linguistic Inquiry and Word Count (LIWC) software, which calculates the relative frequency of syntactic and semantic classes in text samples based on val-

[11] This split rests on the assumption that physicians may share common styles. Thus, the testing data will represent different physicians, but the styles themselves have been captured by the training data so that they can be correctly classified; the same rationale can be applied to image cases. To further investigate the phenomenon and identify the degree of inter- and intra-individual variation in decision style, future work could experiment with holding out particular images and physicians.

[12] In Section 4.1, parameters were tuned for each case of feature combinations in a similar way.

idated, researched dictionaries (Tausczik & Pennebaker, 2010). *Disfluency* features were silent pauses, and the frequency of fillers and nonfluencies as computed by LIWC. *Speech* features are in Table 3.

Besides linguistic features, three additional groups of features were included, with an eye towards application. *Demographic* features were gender and professional status, while *cognitive* features were physician confidence in diagnosis and correctness of the final diagnosis. *Difficulty* features consisted of an expert-assigned rank of diagnostic case difficulty, and the percent of correct diagnoses given by physicians for each image, calculated on the development data only. In an instructional system, a trainee could input a demographic profile, and the system could also collect performance data over time, while also taking into account stored information on case difficulty when available. This information could then be used in modeling of decision style in spoken or written diagnostic narratives.

3.2 Feature Selection

WEKA's CfsSubsetEval, an attribute evaluator, was used for feature selection,[13] using 10-fold cross-validation on the development set only. Features selected by the evaluator in at least 5 of 10 folds were considered best features. The best features from the entire feature set were: *2nd person pronouns, conjunctions, cognitive process, insight, cause, bio*, and *time* words, plus *silent pauses, speech length, time of min. pitch, standard deviation of pitch, time of min. intensity*, and *difficulty: percent correctness/image*.

Feature selection, using the same attribute evaluator, was also performed on only the lexical features, which could be a starting point for analysis of decision-making style in text-only data. The best lexical features[14] included conjunctions, cause, cognitive process, inclusion, exclusion, and perception words. These lexical items seem associated with careful examination and reasoning, which might be more present in analytical decision-making and less present in intuitive decision-making. Some categories, especially inclusion (e.g., *with, and*), exclusion (e.g., *but, either, unless*), and cause words (e.g., *affect, cause, depend, therefore*), seem particularly good representatives of logical reasoning and justification, a key feature of analytical reasoning. But as shown in the next section, when available, speech and disfluency information is useful, and potentially more so than some lexical features.[15]

4 Results and Discussion

Table 4 lists the results for the Random Forest (Breiman, 2001) and Logistic Regression (Cox, 1972) classifiers on the best features (as selected from all features) on the final test set, after training on the development set. These results suggest that decision style can be quantified and classified on a binary scale; the percent error reduction (compared to baseline performance) for both classifiers is substantial.

Classifier	**%Acc**	**%ER**	**Pr**	**Re**
Random Forest	88	76	88	88
Logistic Regression	84	67	84	84
Majority Class Baseline	51	–	–	–

Table 4: Performance on final test set; reduction in error is calculated relative to majority class baseline. Precision and recall are macro-averages of the two classes.

4.1 Feature Combination Exploration

A study of feature combinations was performed on the final test set with Random Forest (Table 5) to explore the contribution of each feature type towards automatic annotation. The best performance was achieved after applying feature selection on all features. Lexical and disfluency features were useful for determining decision style, and the best linguistic features (chosen with feature selection) were slightly more useful. These latter feature types improve on the performance achieved when considering only

[13] With BestFirst search method.

[14] Best lexical features were: function words, singular pronouns, prepositions, conjunctions, quantifiers, and cognitive process, cause, discrepancy, tentative, inclusion, exclusion, perception, see, bio, motion, time, and assent words.

[15] Feature selection was also performed only on the linguistic (lexical, speech, and disfluency) features as a group. The best features of these types were: second personal pronouns, conjunctions, cognitive process, insight, cause, bio, and time words; silent pauses; and speech length, time of minimum pitch, standard deviation of pitch, and time of minimum intensity. They could represent a starting for point for analyzing speech data not enhanced by additional speaker and task information.

speech length and silent pauses, which were apparent characteristics to the human annotators and among the best features (see Section 3.2.).

Demographic features improved somewhat over the baseline, indicating an association between gender, professional status, and decision-making, and adding cognitive features increased performance. Importantly, overall these findings hint at linguistic markers as key indicators of decision style.

Features	**Accuracy**
All*	88
All	85
(Lexical + Speech + Disfluency)*	86
Lexical + Speech + Disfluency	84
Lexical + Disfluency	84
Only speech length and silent pauses	81
Disfluency	79
Lexical	77
Demographic + Cognitive	68
Demographic	64
Majority Class Baseline	51

Table 5: Performance on final test set. Star (*) indicates the use of feature selection (see Section 3.2.)

4.2 Limitations

In this study, doctors diagnosed solely on the basis of visual information (e.g., without tests or follow-up), so their speech may reflect only part of the clinical reasoning process. In addition, most decision style ratings on the 4-point scale were in the distribution center (Figure 2), so the binary labels used in the study only partially reflect purely intuitive or purely analytical reasoning. However, since clinician reasoning in the current dataset can be reliably measured by human and computational classification, linguistic features of decision style must be present. Finally, the LIWC software used for lexical features matches surface strings rather than senses; future work might operate on the sense rather than token level.

5 Related Work

Lauri et al. (2001) asked nurses in five countries to rate statements representative of intuitive or analytical decision-making on a 5-point scale. They found that reasoning varies with context and that styles in the middle of the cognitive continuum predominate. In this work, annotation ratings were prevalent in the middle of the spectrum. Thus, both studies endorse that most decision-making occurs in the central part of the continuum (Hamm, 1988; Hammond, 1981). Womack et al. (2012) proposed that silent pauses in physician narration may indicate cognitive processing. Here, silent pauses were also important, perhaps because analytical decision-making may recruit more cognitive resources than intuitive decision-making.

6 Conclusion

This work suggests that decision style is revealed in language use, in line with claims that linguistic data reflect speakers' cognitive processes (Pennebaker & King, 1999; Tausczik & Pennebaker, 2010). Theoretically, the study adds validity to the dual process and cognitive continuum theories. Methodologically, it articulates a method of transitioning from manual to automatic annotation of fuzzy semantic phenomena, including label adjudication and data selection for computational modeling. Future work may investigate modeling of the 4-point decision scale, as well as whether particular variables, such as difficulty or expertise, mediate the relationship between diagnostic correctness and decision style.

Practically, automatic detection of decision style is useful for both clinical educational systems and mission-critical environments. Clinical instructional systems can assess whether trainees are using the appropriate style for a particular task (Hammond, 1981), and they can help users determine and attend to their own decision styles, towards improving diagnostic skill (Norman, 2009). Finally, in mission-critical environments, linguistic markers of decision-making style may be used to determine the optimal modes of reasoning for a particular task in high-stakes human factors domains.

Acknowledgements

This work was supported by a COLA Faculty Development grant, Xerox award, and NIH award R21 LM01002901. Many thanks to the annotators and reviewers. This content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Allinson, C. W., & Hayes, J. (1996). The cognitive style index: A measure of intuition-analysis for organizational research. *Journal of Management Studies, 33*(1), 119-135.

Alm, C. O. (2011, June). Subjective natural language problems: Motivations, applications, characterizations, and implications. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short papers-Volume 2* (pp. 107-112). Association for Computational Linguistics.

Altman, D. (1991). *Practical statistics for medical research*. London: Chapman and Hall.

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *American Journal of Medicine, 121*, S2-S23.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Cader, R., Campbell, S., & Watson, D. (2005). Cognitive continuum theory in nursing decision-making. *Journal of Advanced Nursing, 49*(4), 397-405.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213-220.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B, 34*(2), 187-220.

Croskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine, 78*, 775-780.

Croskerry, P., & Norman, G. (2008). Overconfidence in clinical decision making. *The American Journal of Medicine, 121*(5), S24-S29.

Evans, J. (1989). *Bias in human reasoning: Causes and consequences.* Hillsdale, NJ: Erlbaum.

Evans, J. (2008). Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology, 59*, 255-278.

Graber, M. (2009). Educational strategies to reduce diagnostic error: Can you teach this stuff? *Advances in Health Sciences Education, 14*, 63-69.

Graber, M. L., Kissam, S., Payne, V. L., Meyer, A. N., Sorensen, A., Lenfestey, N., ... & Singh, H. (2012). Cognitive interventions to reduce diagnostic error: A narrative review. *BMJ Quality & Safety, 2*(7), 535-557.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10-18.

Hamm, R. M. (1988). Clinical intuition and clinical analysis: Expertise and the cognitive continuum. In J. Dowie & A.S. Elstein (Eds.), *Professional judgment: A reader in clinical decision making* (pp. 78-105). Cambridge, England: Cambridge University Press.

Hammond, K. R. (1981). *Principles of organization in intuitive and analytical cognition (Report #231).* Boulder, CO: University of Colorado, Center for Research on Judgment & Policy.

Hammond, K. R. (1996). *Human judgement and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice.* New York, NY: Oxford University Press.

Hochberg, L., Alm, C. O., Rantanen, E. M., DeLong, C.M., & Haake, A. (2014). Decision style in a clinical reasoning corpus. In *Proceedings of the BioNLP Workshop* (pp. 83-87). Baltimore, MD: Association for Computational Linguistics.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics of intuitive judgment: Extensions and applications* (pp. 49-81). New York, NY: Cambridge University Press.

Lauri, S., Salanterä, S., Chalmers, K., Ekman, S. L., Kim, H. S., Käppeli, S., & MacLeod, M. (2001). An exploratory study of clinical decision-making in five countries. *Journal of Nursing Scholarship, 33*(1), 83-90.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Norman, G. (2009). Dual processing and diagnostic errors. *Advances in Health Sciences Education, 14*(1), 37-49.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*(6), 1296-1312.

Sjöberg, L. (2003). Intuitive vs. analytical decision making: Which is preferred? *Scandinavian Journal of Management, 19*(1), 17-29.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24-54.

Womack, K., McCoy, W., Alm, C. O., Calvelli, C., Pelz, J. B., Shi, P., & Haake, A. (2012, July). Disfluencies as extra-propositional indicators of cognitive processing. *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics* (pp. 1-9). Association for Computational Linguistics.