StageLens

MetaIntegral Project—Research Results Report[1]

Exploratory Research on Automated Text Analysis for Assessing Group Developmental Level

Tom Murray
Terri O'Fallon

(in consultation with Suzanne Cook-Greuter)

March 10, 2015

Contact: tommurray.us@gmail.com
Website: www.stagelens.org

[1] Acknowledgements: We would like the thank the MetaIntegral Foundation for its generous support of this work. Also, thanks to Dr. Keith Otis for statistics and data mining consultation, and to Matthew Rosenblum for data processing assistance. Thanks to Geoff Fitch and Venita Ramirez for recoding Delta data, mentioned below.

Preface

This report summarizes research work done thanks to funding from the MetaIntegral Foundation, through a project titled "Exploratory Research on Automated Text Analysis for Assessing Group Developmental Level." The purpose of this project, which ran from July 1, 2014 through March 1, 2015, was to do exploratory research to test the feasibility of using state-of-the-art text analysis, machine learning, and data-mining methods to recognize developmental level, based on Terri O'Fallon's new StAGES scoring method for scoring the 36-sentence completion MAP ego/leadership development instrument.

Introduction and Motivation

Can we train computers to classify developmental levels from text? Text analysis and artificial intelligence (machine learning or data mining) methods have advanced notably in recent years, to a point where it is feasible to imagine automated methods that approximately classify developmental in ego development, reflective reasoning, leadership maturity, and social metacognition. We do not expect that automated methods will be as reliable as human scoring, but there are many potential applications for quick inexpensive scoring that have a larger margin of error than human scoring—particularly in taking averages across groups of people. Here are some example of the applications that would be possible:

(1) Imagine: easily scoring the pre and post developmental levels of 500 people who went through an educational program.

(2) Imagine being able to inform the design of an organizational change plan through a scan of a group's Twitter or blog comments to get an average estimate of developmental level.

(3) Imagine planning a conflict resolution process between two groups from an analysis of their introductory remarks, to discover that one has an Expert center of gravity and the other has a Pluralist center of gravity.

(4) The medical profession is plagued by patient's lack of compliance with the advice or orders of medical professionals. Much of this might be because patient's ability to understand and the values that motivate them, are strongly determined by developmental factors; while doctors may be blind to this and communicate instruction in ways that are not very effective. Imagine having easy inexpensive ways to help doctors estimate the type of information and motivational methods that are more likely to work, individualized to each patient.

This work builds upon: (1) Tom Murray's years of academic research in Artificial Intelligence and, more recently, in state-of-the-art text analysis to identify social deliberative skills in online dialogue; (2) Terri O'Fallon's recent extension (or alternate re-modeling) of Susanne Cook-Greuter's leadership/ego development scoring system (MAP), that scores through an AQAL lens to simplify and add deep structure to the original method. This makes it more amenable to atomization; (3) the expertise of Susanne Cook-Greuter, who is consulting closely with us.

Hand-scoring of sentence completion tests is very labor intensive. Automated analysis would allow for a rough analysis of group developmental level when such scoring is prohibitive by cost or logistics. We also hypothesize that such automatic scoring could be applied to assess that developmental level of any text, not only the MAP sentence completion instrument. This would open up a huge new space for integral and developmental large group practices/interventions and also for research projects studying developmental aspects of psychological, social, or political phenomena. This small MetaIntegral Foundation R&D grant allows for first steps toward what promises to be an extended research agenda for multiple scholars over many years.

The StAGES Model

We assume that the reader is familiar with the ego/leadership maturity developmental model, and the 36-item sentence completion instrument used by Cook-Greuter, which in turn was based on research by Loevinger (Cook-Greuter, 200, 2005; Loevinger, 1976). The StAGES model, recently developed by O'Fallon, proposes a new model for understanding the underlying structure of the progression of stages (O'Fallon, 2011). O'Fallon has also developed a new scoring system for scoring sentence stems based on the new model. Early assessments indicate that the StAGES scoring method produces stage assessment results that correlate with Cook-Greuter's method with sufficient statistical significance (O'Fallon, 2015; in press; and further tests of psychometric properties of the new system are ongoing).

Figure 1 illustrates the levels in the StAGES model. It includes 12 levels which are divided in three Tiers: Gross, Subtle, and Causal. 'Person Perspectives" associated with each stage are shown in the Figure (for example, Expert is Early First Person and Achiever is Late Third Person perspective). The StAGES model is compatible with Cook-Greuter's model (which I will call the MAP model), but has some additions. The MAP Diplomat level is split into Delta and Diplomat; and Cook-Greuter's highest level is separated into three levels (Transpersonal, Universal, Illumined).

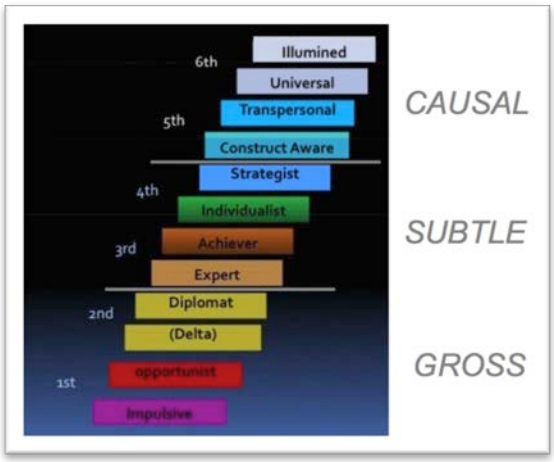

Figure 1: 12 Stages with 6 Person-perspectives

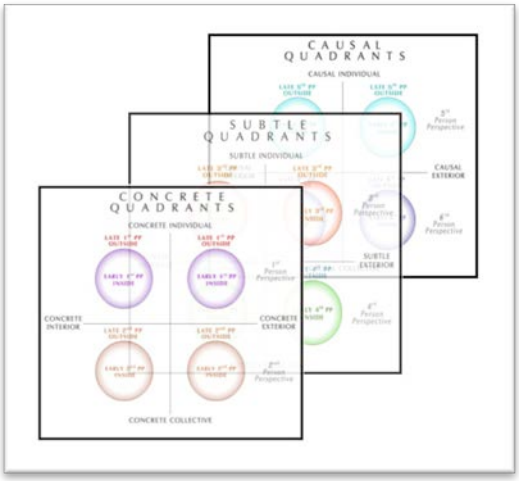

Figure 2: Diagram of Stage Tiers vs Quadrants

Figure 2 illustrates how the StAGES model is based on Wilber's 4-quadrant/8-zone model, with stage levels progressing in the same way through the quadrant map repeated for each of the three tiers. The progression is further illustrated in Figure 3. Within a given tier, the Person Perspectives progress from Individual to Collective oriented. Within each Person Perspective stages progress from Inside (passive-had-by, related to 1st person) to Outside (active-having, related to third person). Thus, for example, the Achiever stage can be characterized by three indices: Subtle Tier, Singular, Outside. Each of the 12 levels can be described in this way using only three attributes. Figure 3 also shows that the Interior/Exterior aspect of the quadrant model translates in to the difference between Early and Late within each stage. We will ignore the Early/Late (Exterior vs Interior) designation in what follows, as we will only be interested in identifying stages, not sub-stages. For explanations and arguments for *why* the stages should progress in this particular way, see O'Fallon (2011, 2013).

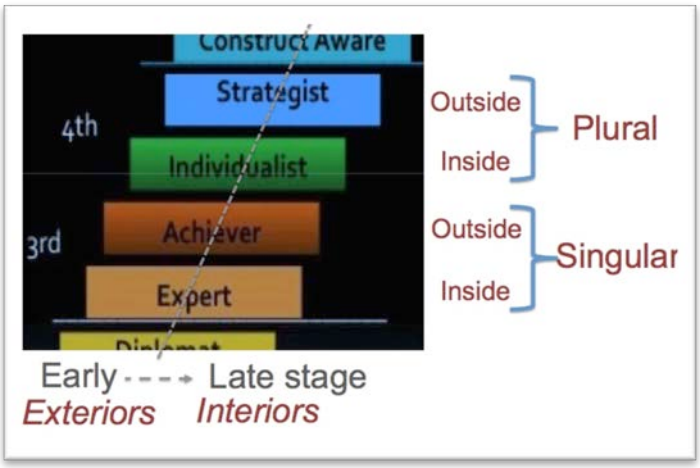

Figure 3: Illustration of Quadrant components within a Tier

Overview of Computational Modeling Goals

The goal of this project is to explore whether we can build computational models that take the text of a sentence completion item as input and predict the stage as an output, with reasonable accuracy as compared with human coding. A corollary goal is for a model to take an individual's 36-sentence completion items as input and return the overall stage score for that person, with reasonable accuracy as compared with human coding. As explained in the Motivation, we do not expect the automated analysis to be as accurate as human coding, nor does it need to be for it to be useful for some applications.

Text (or document) classification is a maturing field within computer science that has overlap with Data Mining, Text Anlytics, and Machine Learning fields (Graesser et al., 2010; Rosé et al 2008; Stolcke et al. 2000; Witten & Frank, 2005; Cherkassky, 2013). The type of classification we are doing is called "supervised learning" because we are training a model with examples with known classification (stage level). Supervised text classification methods are being used for many applications, including filtering email spam, classifying the genre or readability of an article, automatically grading school essays, and identifying emotional tone ("sentiment analysis"). The modeling task is made simpler the fewer the categories ("labels") there are. For example, in email spam filtering the text either is or is not spam, and the modeling algorithm is fed positive and negative examples of email spam in an attempt to train it. In some of our work we are trying to model 12 Stages, which could be quite difficult. However, modeling whether a sentence is Individual or Collective might be easier, at least in terms of the model complexity.

Classification models (classifiers) are built by "training" them with input examples that have known correct answers ("label"), as illustrated in Figure 4 (a). Once a model is built is it used to "predict" the classification (label) for future inputs (b). Part of the training usually involves "testing" the model or verifying it using a set of data with known correct answers that was not used to train it. Note that part of the process is to extract "features" from the input data. In text analysis there are many approaches to extracting features. For example, whether or not the text has a first-person pronoun might be one feature. In text classification the model can use of hundreds or even thousands of such features. Features separate from the text might be used, for example the age of the person (not always available, and we did not use this feature), which might contribute to predicting the stage. The model is essentially like a huge equation or decision tree that takes in many features as input and outputs a label to classify the input.

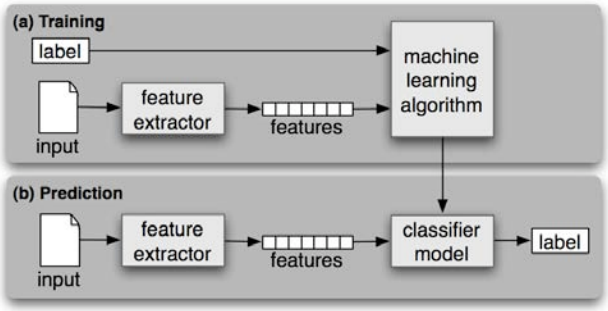

Figure 4: Machine Learning

Until recently, scoring the MAP sentence completion instrument was done by scorers trained using a scoring manual that is approximately 300 pages long, with details and examples for each of the 36 sentence stems. One might expect that training a computer to, in essence "know" this manual might be exceedingly difficult. With the advent of the StAGES model, the vision of modeling the scoring process was dramatically changed, and this is what inspired the StageLense project. From the perspective of StAGES, scoring a sentence completion became, at least conceptually, a process of answering just three questions for the three dimensions of the model (leaving out the forth interior/exterior dimension): What is the tier (G/S/C)? Is it speaking to an individual (singular) or collective (plural) object-space? Is it an Inside or Outside perspective on that object space? Now, these questions are not easy to answer in many instances, but (assuming that StAGES is representing essentially the same developmental trajectory that MAP was, for which there is strongly suggestive, but still inconclusive evidence) the prospect of "teaching" a computer to score the sentences now seemed more possible.[2]

Because of this reconceptualization, the StAGES scoring manual is about 1/10th the length of the MAP scoring manual. Answering these three questions is still a difficult task requiring significant training and intelligence. A description of what these three dimensions mean is outside the scope of this papers (see O'Fallon 2011), but note that the questions are not all simple grammatical categories.

Figure 5 illustrates the three dimensions G/S/C, S/P, I/O[3] associated with each of the 12 levels. It also shows two additional labels. The first is perspective, mentioned above. The second is Complexity, which is a scalar that goes from 1 to 4 within each tier. O'Fallon has noted repeating patterns in this sense also. Though this idea has not been fully worked out, it appears that there are interesting relationships between Complexity and the orders of complexity within the developmental models of Fisher and Commons (as used by Dawson and Stein).[4] Thus there are, if effect, 6 types of patterns that O'Fallon has described in her theory, that are structures as:

- Stage: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
- Perspective: 1, 2, 3, 4, 5, 6
- Tier: G, S, C
- Complexity: 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4
- Person: S, P, S, P, S, P

[2] The algorithm for going from the scores of each of the individual 36 sentence completions to the full score for an individual is identical for StAges and MAP. It is not a simple averaging. Both algorithms weight higher scores more, and both consider other actors in the final analysis, such as the spread of scores.

[3] I/O: inside/outside; or active-having/passive-had-by; also related to 1st/3rd person subjective/objective stance.

[4] Complexity within a StAges tier progresses in the complexity of the relationship among objects of the given type (gross, subtle, or causal objects): from simple and *receptive*, to *either-or and active* stance, to *both-and reciprocal* stance, to *interpenetrative* active stance. This seems to parallel the progression from single set to mapping to system to system of systems within each "Tier" (order of abstraction) in the Hierarchical Complexity model (Dawson, 2004; Dawson & Stein, 2011; Fisher, 1908; Commons & Richards, 1984).

- Active/Passive stance: I, O, I, O, I, O, I, O, I, O, I, O

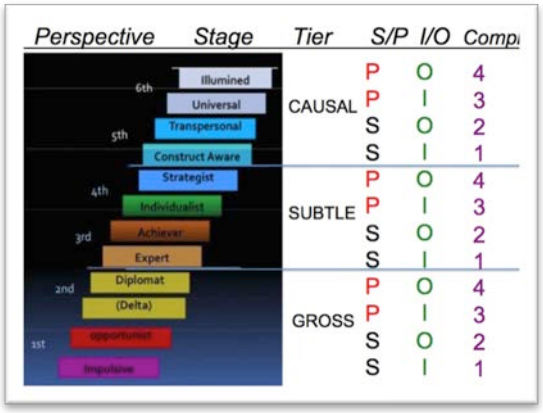

Figure 5: Classification Dimensions

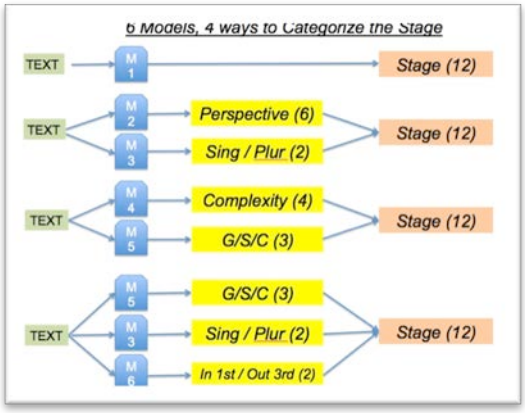

Figure 6: Stage Model Alternatives

One could try to build computer models to predict any of these six categorizations. Figure 6 illustrates several paths to classifying the final stage. Note that it is not simply the "text" that is the input of a model (M in the Figure), but features that are extracted from the text (as in Figure 4). The top of the Figure shows that we can build a model that tried to predict Stage directly. Alternatively, if we had models that accurately predicted G/S/C, S/P, and I/O, it is trivial to then calculate the stage. The Figure shows that there are four such paths to predicting stage based on text. Our goal is to experiment with building models for all six of these categorization possibilities.

Text Features

There are countless ways one could specify the features of text. One method used in AI (artificial intelligence) text classification is called the "bag of words" method, in which the features used to train the model are simply a list of the words in the text (perhaps modified by the word frequency). An algorithm can look for patterns in singles words or word-sequences, though doing so is computationally more time consuming (i.e. "red car" can be considered one of the features). The part of speech (POS) of each word (e.g. noun, adverb, gerund...) can also be used as features. We are experimenting with bag-of-word methods, but are focusing on extracting more information-rich features from the text and letting the data mining algorithms find patterns within these features. One possible way to do this would be to comb the scoring manuals (StAGES or MAP) for specific word patterns, and use these as features (also adding any text features mentioned by our experts Cook-Greuter and O'Fallon). For example, MAP scoring mentions the use of either/or vs. both-and types of language as an indicator. The way people use pronouns (I, you, we, they) is also mentioned; as are words related to right/wrong and should/shouldn't. This would, in effect, be like building an "expert system" of rules that tries to capture the expertise represented in the scoring manuals. Our research team would need to go through the manuals to identify every possible parameter or feature (or the most promising of them), then it would be left to the machine learning algorithm to build the exact "rules" that specify, in a very complex way, what combination (or proportion) of features predicts the category (stage, perspective, etc.).

We did not use this method for several reasons, though we do plan additional research in this direction. Though the examples given above seem straightforward (i.e. searching for "should/shouldn't" or pronouns) the vast majority of the features or "rules" implied in the scoring manuals are not easily translated into computational terms—they are intended for human

understanding. For example, development tends to progresses from vague to specific words, and from ego-centric to ethno-centric; one's relational space progresses from narcissistic to like-situated to like-minded to like-principled; tone of voice progresses from more helpless or hostile to earnest, to pluralistic and tolerant; from immediate to short term to longer term anticipation. Even identifying what words imply Subtle as opposed to Gross objects is deceptively difficult (not even to mention Causal objects). These attributes are extremely difficult to translate into text properties that are simple enough for the computer to assess.

Thus we are using another approach to feature specification and extraction. There already exists within the text analysis field many classification methods and metrics. Although these metrics were not designed to measure ego/leadership development, we hope that our data mining algorithms will find patterns or correlations within these features that are predictive of stage level, perspective, tier, person, complexity, or orientation. We have access to text analysis methods that produce over 200 such metrics. Thus the process of feature extraction creates over 200 numbers (metrics or features) for each sentence completion item. These types of features have been successfully used in the past to create models for psychologically relevant attributes such as emotional tone, deception, personality style, and conceptual complexity, so we expected they may be useful for developmental level as well. (Remember that all models have uncertainty and error, and that in some studies, for example a 70% success rate in classifying text items is considered an achievement—yet 30% of the predictions are wrong.) These metrics are the features that we use to train our models. Example metrics include the following concepts, expressed as numeric measurements:

- Number of nouns, verbs, adjectives, etc.
- Lexical diversity
- First, second, and third person pronouns
- Number of modifiers used
- Intentional verb count
- Mean sentence length
- Present and past tense usage
- Negative vs. Positive emotion words
- Word concreteness vs. abstraction
- Words associated with certainty, causation, motion
- Syntactic and semantic similarity between adjacent sentences
- Approximate reading grade level for the most sophisticated words used
- Word familiarity (or uniqueness)
- Word polysemy (number of meanings or senses in the dictionary definition).

We could have combed the list of over 200 text metrics to specify exactly which ones we expected should be correlated (or involved) in each of the six types of classification, and we could have limited the metrics (text features) to these. But this research is more exploratory than hypothesis-driven in this regard—we were curious to see which of all of these metrics were related to the six classification schemes. Also, we did not completely understand the implications of many of the metrics (e.g. temporal cohesion, left-embeddedness), but it was still possible that such features would play an important role in classification. Thus we included all metrics that seemed applicable, resulting in 155 metrics (for example we excluded those that were more appropriate to the analysis of multi-paragraph text items). This allowed the modeling to use as much information as we could make available, and also would allow us to discover patterns or correlations that were unexpected, including properties of the text that were not mentioned in the scoring manuals but were nevertheless found to be relevant.

We also included the stem-number (1-36) as a learning feature, allowing the machine learning methods the possibility of finding different patterns depending on the stem (which vary in subject matter from, e.g. family relationships, work or professional domains, self-reflection, emotional reflection). Though we may try this in the future, we did not include age or gender information within the features.

Data Description and Feature Statistics

Our original plan was to use data from studies done through O'Fallon and Pacific Integral, with Cook-Greuter consulting on methods and outcomes, and providing some data for comparative purposes. Due to an extended illness within O'Fallon's family, Terri was unavailable to organize and pre-process the data necessary for this project during the critical months. Cook-Greuter therefore supplied data (sample sentence completions with correct scoring) from her scoring manuals and from data sets previously released for research purposes. This allowed us to make significant progress, but the results have some limitations, as the scoring systems for MAP vs StAGES are different.

Here are some implications of the change in data sources:

- One challenge was that the StAGES system has an additional level, splitting the Diplomat level in MAP into Delta and Diplomat. (This extra stage was originally included in the model used by Loevinger and Cool-Greuter, but was combined with Diplomat for a number of practical reasons, including rarity of data in that category.) We got generous volunteer efforts from Pacific Integral staff to re-score the MAP Diplomat scores as either Diplomat or Delta.
- One result of the change from StAGES-scored to MAP-scored data is that the models that we build for predicting the StAGES categories (Gross/Subtle/Causal; Individual/Collective; and Inside/Outside) are less valid; while the models for predicting stage level are relatively more valid.
- Another major limitation of having used the MAP data is that most of the data we have is from examples from scoring manuals, which are organized by sentence stem (e.g. for a given stem, examples are given for all levels). We have relatively little data on a per-person level, and thus all of our modeling is on the per-stem level, not per-person level. I.E. the accuracies and other metrics are for the ability to predict the scoring of sentences, not the ability of the model to predict the overall scoring for a person.

In the future we plan to redo our analysis using O'Fallon's StAGES-scored data, but all results presented here are based on the MAP data set. The silver lining to this change is that we are gaining insight into the MAP data set in more depth than was planned. Also, this sets us up for more thorough comparison of MAP vs StAGES scoring once we get StAGES-scored data.

Descriptive statistics: We are working with example sentences from approximately 120 pages of the MAP scoring manual, which covers all MAP levels for 12 of the 36 sentence stems in the MAP instrument, yielding **5388 responses**. In addition we have data from 11 individuals (x 36 stems = 396 responses). Thus our data set is comprised of 5784 sentence items and their associated stage rating. All of these example text items, along with the correct stage score, were laboriously culled from raw data text documents (both scoring manual pages and MAP scoring result report sheets) in which formatting variations etc. prevented automated extraction of the responses. Thus over 90% of our data is from the manual pages.

A FileMaker database solution was developed to store and preprocess the data. This step included parsing out the labels associated with all 6 classification types (based simply on the stage, e.g. an Expert stage sentence is also has Inside and Singular classification), and other preprocessing that, for example, removed non-standard characters from the text.

Text analysis for feature extraction. The next step was to process each of the 5388 responses (we will sometimes just call them "sentences" even though some responses are multi-sentence) through the text analysis methods to produce 155 metrics for each response. We then produced graphical visualizations of the distributions and relationships between each metric and the six classifications (this was done through programming in the R statistics language). Figure 7 illustrates these charts. To the left are histogram charts. These show the overall frequency distribution each of the 155 metrics over the entire data set. The integers in the corner indicate the number of unique values. Y values are frequency within a bin, and X values show the range of values possible for each metric (which differs per metric). For example the central left graph shows a roughly normal (bell shaped) distribution; the lower left shows a U-shaped distribution, and the upper right shows a metric for which all most all of the data has the same value (which may indicate it is not useful to include).

In the center of the Figure are examples of contour plots for all metrics vs stage level. You can see that some of the metrics show definite patterns from early (left) to later (right) stages, while others show little difference. This gave us some indication that many of the metrics were indeed relevant. To the right in Figure are box plots, which are used to illustrate relationships between the 155 metrics and each of the categories (G/S/C is shown in the Figure). Again, these visualizations gave an indication that some metrics were indeed relevant to the task of classification. The graphs are also useful later in the analysis process. For example if a metric proves to be highly correlated with stage, the contour plot gives much more information about the nature of that relationship than a single correlation number. For example, the lower right contour plot indicates that this metric has a similar low value for lower stages, then jumps up to flatten out at a higher level for higher stages. Deeper investigations can ask "what is going on at that jump?"

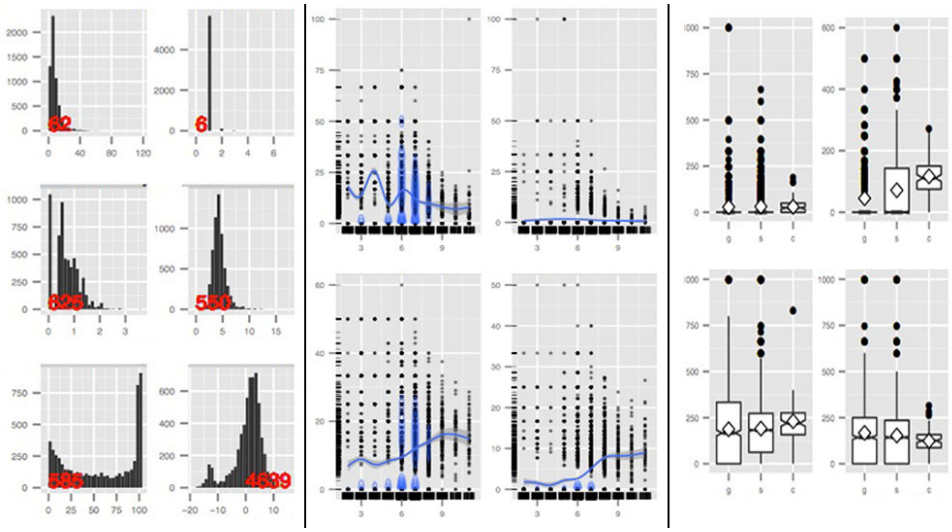

Figure 7: Data Plots: A: Histograms B: Contour Plots, C: Box Plots

In addition to the charts shown above, we also produced correlation tables showing the correlations between the 155 metrics and stage (and perspective and complexity), and also showing the correlations between each pair of metrics.

Model Research and Construction

The next task was to use a variety of machine learning algorithms to experiment with building models for each of the 6 types of classifications. We used a variety of methods and tools to do this, but they have the basic methodology in common. After the text is processed to

produce features (155 of them for almost 6000 data points) these features, along with the "correct" classification, are fed into a machine learning algorithm, of which there are many variations. The output is a model that can be used to predict future exemplars, plus various measurements that characterize how the model performed in relation to its test data.

The classic train/test method is to split the data into a training set and a test set, because it is like cheating to test the model on the same data used to train it. This is largely due to the fact that models can "over-fit" which means that they are tuned to match the training set but don't generalize well to new cases. The best models are those that neither over-fit nor under-fit. There are many ways to characterize the performance of a model, including its accuracy, precision, and recall. These variations get at the fact that the errors made by models can lean more toward false-positives or false-negatives, and there are various ways to describe the error patterns. An additional useful metric is the Cohen's Kappa statistic, which is conceptually similar to accuracy but takes into account the fact that random guessing can produce correct classifications (it measures how much better than a random guess the model does). Perhaps the most informative picture of a model's performance is though its "confusion matrix."

Figure 8 shows a confusion matrix for a particular model predicting G/S/C, showing the actual vs. predicted label or category. For example it guessed 4042 of the Subtle examples correctly, but misclassified 654 of the Gross examples as Subtle. The goal is to get high numbers on the diagonal—this model showed an 83% overall accuracy (which looks at that diagonal). In working to understand and improve this model, one would further investigate how the model was performing in any non-diagonal cell with an unacceptably high value. For example, what is it about those 654 Gross examples that makes them classify as Subtle? Investigating this can lead to tweaks to the model that improve its performance.

Act_Pred	g	s	c	rowsums
g	662	654	0	1316
s	287	4042	18	4347
c	0	68	53	121
Sum_or_ave	949	4764	71	5784

Figure 8: Confusion Matrix for G/S/C

One problem with the classic method of separating test data from training data is that you cannot be sure that your test data is a perfect representation of the training data—maybe there are systematic differences. Most modern machine learning methods use another train/test method, called cross-validation. We used 10-fold cross validation (the standard method) which randomly removes 1/10th of the data to use as a test set, then builds a model and tests it as in the classical method. But then repeats this process 10 times, each time removing a different 10% of the data. Then in the end it combines (averages) all 10 models to produce one that is a better overall model than any of the 10.

There are a plethora of machine learning algorithms. The ones we have experimented with thus far include: SVM (support vector machine), Naïve Bayes, Decision Trees, Logistic Regression. Each of these have parameters that can be tuned (and numerous variations on feature selection, pruning parameters, etc.). Our process thus far has been one of defining a set of experiments that define the parameters, building the models, checking the performance of the models, and then iterating based on what we have learned. Because (1) the "search space" of variations is so huge, (2) machine learning is still a "black art" that must be approached with some guidelines but much trial and errors, and (3) in our particular case there have been no clear winners across categories of machine learning methodologies—as of this report we are still in early stages of exploring this territory. The table below shows an example of the results of some experimental runs. In these series we are recording the processing time for the modeling software, the accuracy, and the kappa statistic (confusion matrices are also saved). These are sample results

meant to illustrate the methodology, and are not necessarily representative of best or average results.

Experiment ID	Time (min)	Accuracy	Kappa
b01_cgredu-stagesx2Nom-12gram-nostem_bayes	0.25	0.4697	0.2853
b03_cgredu-stagesx2Nom-12gram-nostem_svm	0.25	0.4303	0.2474
b04_cgredu-stagesx2Nom-12gram-nostem_dtree	15	0.4481	0.2338
b05_cgredu-stagesx2Nom-12gram-stem_bayes	0.25	0.4222	0.2574
b37_cgredu-inout-12gram-stem_svm	0.25	0.5915	0.1831
b38_cgredu-inout--featr-nostem_logl2	0.5	0.5845	0.1691
f01_cgres-stagesx2Nom-featr_bayes	0.25	0.2513	0.157
f04_cgres-stagesx2Nom--featr_logl2	0.5	0.5432	0.3623
f05_cgres-gsc-12gram_bayes	0.25	0.6982	0.3526
f07_cgres-gsc-12gram_logl2	0.25	0.9112	0.6888
f26_cgrres-st2-stagesx2Num_dtree	0.25	0.7972	0.6252
f27_cgrres-st3-stagesx2Nom_dtree	0.25	0.5625	0.4033

Based on this illustration Table, it can be seen that we are experimenting with the following parameters in many (not nearly all) combinations:

- Whether the model is for stage, G/S/C, I/O, S/P, Perspective, or Complexity
- Whether the label is treated as Nominal (categorical) or Numeric (for stage, perspective, complexity only)
- Whether the text includes the stem and response, or the response only
- Whether the data has been "resampled" ("cgrres") to boost the percentage of low and high stage examples, which are underrepresented in our data set.
- The type of learning algorithm (decision tree, Bayes, SVM, or logistic)
- Whether the bag-of-words ("12gram") or metrics were used as features.
- We have also run some tests (indicated by "st2" "st3" in the Figure) that build separate models for each stem (i.e. 36 models all having the same configuration but specializing in one stem)

Preliminary Results from Modeling Experiments

We have run about 150 such trials to date. We are still very much in a stage of breadth-first exploration of this search space, and looking to the data mining literature for clues as to best practices and hints. We are not finding strong patterns yet that allow us to reject entire sections of possibilities. For example, though it is looking like logistic regression is the winning search method, it often is only slightly better and the instances where it is worse do not seem to have a consistent pattern as of yet. Also, there are some surprising anomalies in the processing times, which can range from a few seconds to over an hour, with no clear predictor on what types of experiments will take longer times (based on "Pivot-Table" analysis of the results).

Some of our conclusions so far are:

- Many of the metrics (features) have moderate to strong correlations (.25 to .45) vs. the outcome measures, and have very low p-values (less than .001) associated with those correlations.
- Adding the sentence stem to the response does not improve performance

- Numeric representations yield better results than nominal, when applicable (for the same categories; this was expected)
- Text-analysis features (155 of them) usually outperform bag-of-words methods, even when bag-of-words methods include part-of-speech tagging.
- Logistic Regression (L2 type) is overall the most powerful learning method; though for certain situations other methods seem strong (thus we are not ruling any out yet).
- A proportional resampling method that samples sentences according to the stage proportions seen in average adult populations produces significant performance improvements.
- Creating separate models for each stem improves performance.

We can celebrate that some of the models have encouragingly high accuracy levels given our early exploratory phase of this research. For example we have built models with as high at 85% accuracy (Kappa .68) for stage; 91% accuracy for G/S/C (Kappa .69), 85% (Kappa .61) for S/P, and 78% accuracy (Kappa .55) for Perspective. However, we interpret these results with caution as we move forward. Our models for I/O, and Complexity are not yet as strong as hoped for. What is also encouraging is (1) we expect improvement when we begin working with the StAGES-scored data vs. the MAP scored data; and (2) one can expect a significant improvement in performance statistics when move from predicting individual stems to predicting the full individual scores (over 36 stems; which, again, will be possible with the StAGES-scored data).

In addition to the modeling experiments described above, we have been evaluating the correlation tables to gain deeper understanding of how the metrics are contributing to the models (we are also experimenting with stepped linear regression for the same types of insights in predicting stage). The list below shows some of the metrics that are moderately correlated with stage score (or weakly correlated with very low p-values, i.e. high significance).[5]

Positive correlations:
- Reading level (words matched to grade level when introduced)
- Average word length
- Num. words before first verb (est. working memory load)
- Sentence length
- Number of logical relationships
- Rare words
- Abstract words
- Conjunctions

Negative correlations:
- Narrativity (vs. formality of language)
- Negative emotion
- First person pronouns

These metric correlations are compatible with current understanding of ego/leadership development, and most of them are explicitly or implicitly referred to in the MAP scoring system. We are only beginning to assess how these and other correlated text properties manifest at various developmental levels.

Conclusions with Ethical and Practical Concerns

[5] All "number of words" measures are normalized to text length, so are essentially percentages.

The results listed above are quite preliminary and tentative. As mentioned, we plan to continue the work and look for more funding to do so. Our broad goals are as follows:

- Discover model methods and build models yielding high accuracy when used in group-averaging contexts (see Introduction).
- Extend this work to build models that might be useful in providing individual stage evaluation, but to be used only in conjunction with human consultation and oversight.
- To use text analysis to more deeply understand the differences, at a linguistic level, between different developmental levels; and understand the constructs of G/S/C, S/P, I/O, and Complexity.
- To extend this work to the analysis of text beyond the sentence completion tasks; i.e. to automatically analyze developmental level of news articles, web pages, blog comments, etc.
- To investigate extending this work on text analysis to other related domains, such as "spiritual intelligence" or collaboration skill.

The broad potential impacts of this project can be summarized using MetaIntegral's Meta-Impact Framework as below:

DEEP IMPACT Transforming Mindsets	CLEAR IMPACT Transforming Performance
• Developmental assessments open us new avenues of personal understanding, self-reflection, and growth. • As in the example of medical compliance above, professionals with access to inexpensive developmental assessments will be educated in the nature and importance of the developmental perspective in various forms of work.	• Developmental assessment can give clear, specific advice to individuals and their support circles as to their growth edges and challenges, which can lead to behavioral changes and increased capacity for action.
WIDE IMPACT **Transforming Relationships**	**HIGH IMPACT** **Transforming Systems**
• Group-level developmental assessments can spur collective dialogue and self reflection; and deepen collective self-understanding and identity. • This work has the potential to motivate many integrally-informed organizations that make use of developmental assessment to come together in discussing the potential positive and negative outcomes of automatic scoring methods in general.	• Changes noted above in mindset, performance, and relationships will inevitably lead to changes in systems, as organizations become more sensitive to developmental factors. • The easy availability of automated developmental assessment will require the creation of additional systemic levels to handle (1) data processing; (2) providing new services to new audiences and stakeholder types.

Below are some concerns and considerations:

1. If developmental scoring of the MAP sentence completion instrument is made easy and inexpensive, both the potential benefits and all the potential downsides of developmental assessments will be multiplied. The potential for cookie-cutter solutions, rash overgeneralizations, and pigeon-holing individuals and groups based on developmental level will be exacerbated, and new issues will emerge. Also accuracy aside, even valid knowledge of people's developmental levels can be used for ethical and unethical (including manipulative) purposes. This potential will be exacerbated if inexpensive methods are in the "wrong hands."

2. Current methods of scoring MAP for individuals includes carefully authored personalized analysis and advice, which helps people understand the ethical uses and the limitations of developmental scoring. Unless the analysis is done for research purposes, the goal of the assessments is clearly for individual learning and self-reflection, rather than for objective labeling. The availability of quick-and-dirty assessment methods may encourage some to bypass getting this type of careful advice. The dangers exist on many levels, including: individuals; organizations evaluating employees, etc.; the assessment and labeling of groups.

3. If, as is expected, the computer models have significantly more margin or error than human scoring, those using and interpreting developmental scores may not understand these limitations, and treat the less-accurate scores as if they were as valid as human-scored results.

4. The business models of organizations like Pacific Integral and Cook-Greuter & Associates depend in part on training people to be scorers—a task that takes significant expertise and requires extensive training and validation. It is possible that if an inexpensive scoring method was developed that trained to scorers would see a decrease in work, and that organization that train scorers would suffer also.

Responses to Above Concerns:

To address the above issues (*if* the technology is proven to be feasible, which is not determined yet) our first implementations will be limited to group-level analysis of sentence completion tests. These are not for individual "consumption," but for group-level uses as illustrated in the Project Description above (examples #1-3). Also, the technology will not be open-sourced in the near future because of the potential misuse. In addition, human checks-and-balances will be included in the loop to verify that automated methods are valid.

However, we cannot prevent other organizations (with deeper pockets) from using the published results of our studies to build their own models and, rightly or wrongly, advertising that they have technologies that will replace human coding methods. Part of our goal will be to try to educate people in general about the limitations involved in this type of assessment; and the most valid and ethical ways to use it.

Also, in regard to concern #4 above: Automated methods are expected to "increase the pie" for developmental assessment rather than create additional competition for the current market. As the cost drops, many more can become introduced to the possibility of getting an assessment, and might be motivated to pay for the type of personalized counseling that comes with human-scored and supported assessments. Also, automated scoring might be done as a first-pass to increase the speed of human scoring.

Inexpensive assessment methods could introduce the concept of development to orders of magnitudes more individuals; who would then be introduced to and may become curious about the nature and possibilities implied by developmental approaches. This could lead to increased interest and business for all of the integrally-informed organizations who support leadership and organizational development, spiritual development, coaching and education, and social and global change organizations.

References

Cherkassky, V. (2013). *Predictive Learning*. VCTextbooks, U. Minnesota.

Commons, M. L. & Richards, F. A. (1984). A general model of stage theory. In M. L. Commons, F. A. Richards & C. Armon (Eds.), *Beyond formal operations: Late adolescent and adult cognitive development*, (pp. 120-141). New York: Praeger.

Cook-Greuter, S. R. (2000). Mature ego development: A gateway to ego transcendence. *J. of Adult Development,* 7(4), 227-240.

Cook-Greuter, S.R. (2005). Ego Development: Nine levels of increasing embrace. Available at www.cook-greuter.com.

Dawson, T. (2004). Assessing intellectual development: Three approaches, one sequence. *Journal of adult development*, *11*(2), 71-85.

Dawson, T. L., & Stein, Z. (2011). We are all learning here: Cycles of research and application in adult development. Hoare, C. (Ed.). (2011). The Oxford handbook of reciprocal adult development and learning. Oxford University Press.

Fischer, K. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review*, 87(6), 477-531.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., Graesser, A., Kamil, P., Moje, E. B., & Afflerbach, P. (2010). Methods of automated text analysis. Handbook of Reading Research, 34-53.

Loevinger, J. (1976) *Ego development*. San Francisco: Jossey-Bass.

O'Fallon, T. (2011). StAGES: growing up is waking up--interpenetrating quadrants, states and structures. Pacific Integral. www.pacificintegral.com

O'Fallon, T. (2013). The Senses: demystifying awakening Paper presented at the Integral theory Conference, 2013, San Fransisco CA.

O'Fallon, T. (2015-in press). Stages. Albany New York: Suny Press.

Rosé, C., Wang, Y. C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., & Fischer, F. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. International journal of computer-supported collaborative learning, 3(3), 237-271.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, J., Bates, R., Jurafsku, D., et al. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics, 26(3), 39– 373.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.