

# MINING OF TEXT AND DISCOURSE

**CI 5330 – Kristina Robertson, Week 12**

# BOT OR NOT?



Watch

Discover

Attend

Participate

About

Search...



Oscar Schwartz:

## Can a computer write poetry?

TEDxYouth@Sydney · 10:56 · Filmed May 2015

15 subtitle languages ?

View interactive transcript



Watch later



Favorite



Download



Rate

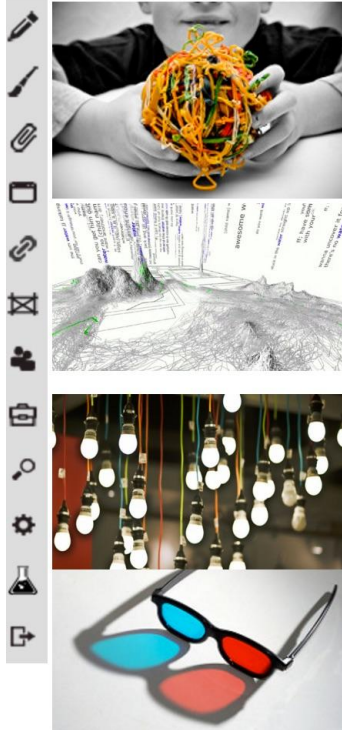
# REFLECTION

WHAT CRITERIA DID YOU USE TO DETERMINE IF THE POEM WAS WRITTEN BY A HUMAN OR A BOT?

WHICH FEATURES OF TEXT MIGHT HAVE BEEN EXTRACTED FROM POEMS AND APPLIED TO COMPOSE POEMS?

# WORD TOOL!

Learning Analytics - Spring'16 Welcome



[bit.ly/la-spring16](http://bit.ly/la-spring16)

- [Week 1 - Introduction](#)
- [Week 2: A Brief Overview](#)
- [Week 3 - Higher order Competencies](#)
- [Week 4 - Learning Theory](#)
- [Week 5 - Hidden Assumptions](#)
- [Week 6 - EDM](#)
- [Week 7 - Cases and Examples](#)
- [Week 8 - Data Hands-on](#)
  
- [Week 10: Learning and Knowledge Growth](#)
- [Week 11: Social Networks](#)
- [Week 12: Mining of Text and Discourse](#)
- [Week 13: Prediction and Intervention](#)
- [Week 14: Applying LA in Practice](#)
  
- [Repository - Community Resources](#)
- [WG Problem Statements](#) [SIGs & WGs Signup Sheet](#)  
Bodong Chen  
1/25/2016, 2:30:01 PM
- [Bios](#) [Random Stuff](#)

# WHAT IS TEXT ANALYSIS?

Word Count

Statistics:

Pages	2
Words	558
Characters (no spaces)	3,038
Characters (with spaces)	3,606
Paragraphs	11
Lines	57

Include footnotes and endnotes

OK

# COHESIVE TEXT ANALYSIS

Coh-Matrix was developed to analyze texts on multiple characteristics and levels of language-discourse. The original inspiration in developing Coh-Matrix was to have an automated metric of **text cohesion** (hence the label Coh-Matrix). Cohesion was of particular interest because comprehension is influenced by some intriguing interactions between text cohesion and the readers' prior knowledge about the topic and their general comprehension skill (McNamara & Kintsch, 1996; Ozuru, Dempsey, & McNamara, 2009).

**Descriptive:** This category will tell you basic information about the text you entered to help you check the summary and make sure that the information makes sense. For example, you will see the number of paragraphs, sentence count, and word count.

**Readability:** Here you can find out how easy or hard your text is to read and understand.

**Syntactic Ease:** Here you will learn about the part-of-speech categories, groups of word phrases, and syntactic structures for sentences. This will include information syntactic complexity, embedded structures, and load on working memory.

**Word Information:** In this category you will find information about grammar categories including nouns, verbs, adjectives, and adverbs of the words used in the text. You will also see information about the function of the words, including prepositions and pronouns.

**Narrativity:** Narrative text tells a story, with characters, events, places, and things that are familiar to the reader. Narrative is closely affiliated with everyday oral conversation.

**Referential cohesion:** High cohesion texts contain words and ideas that overlap across sentences and the entire text, forming threads that connect the explicit textbase.


**Deep cohesion:** Causal, intentional, and other types of connectives help the reader form a more coherent and deeper understanding of the text at the level of the causal situation model.

# LET'S TRY IT - OLD SCHOOL!





# COH-METRIX



## Coh-Metrix

[DOCUMENTATION](#) [WEB TOOL](#) [CONTACT](#)

**What is Coh-Metrix?**

Coh-Metrix is a system for computing computational cohesion and coherence metrics for written and spoken texts. Coh-Metrix allows readers, writers, educators, and researchers to instantly gauge the difficulty of written text for the target audience.

**What is cohesion?**

"Our definition of cohesion consists of characteristics of the explicit text that play some role in helping the reader mentally connect ideas in the text" (Graesser, McNamara, & Louwerse, 2003).

# HOW ARE THE TEXTS DIFFERENT?

## Enter text here:

For five years, I lived on the most beautiful block in the five boroughs of New York City, occupying two full floors of a brownstone on a tree-lined street in downtown Brooklyn.

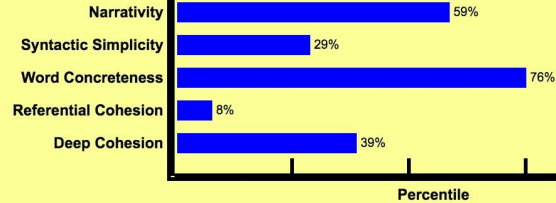
It was the kind of block where the gingko trees turned the evening light gold every October, around the same time families dressed up their stoops with jack-o-lanterns for Halloween. Think the idyllic exterior of the Huxtables' house from the Cosby Show. Think the hood of Spike Lee's She's Gotta Have It – a black creative mecca before the onslaught of gentrification ran every black person out of it.

If my street was picture-book worthy, the house itself was not. I lived in a pre-war building and it chilled me to consider that when it was built, my ancestors were chattel slaves. The facade was crumbling, the iron fence was falling apart. When glass had broken on the front parlor windows, my landlady

Pre-process

Analyze

Clear



Flesch Kincaid Grade Level 7

This text is low in syntactic simplicity which means the sentences may have more clauses and more words before the main verb. Complex syntax is harder to process. It has high word concreteness, which means there are many words that are easier to visualize and comprehend. This text is low in both referential and deep cohesion, suggesting that the reader may have to infer the relationships between sentences and ideas. If the reader has insufficient prior knowledge, these gaps can be challenging.

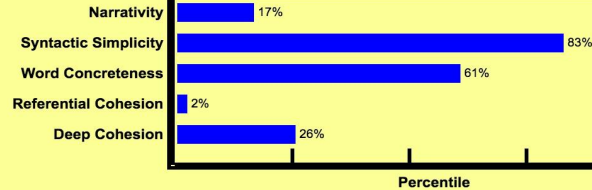
## Enter text here:

was not so wonderful. Near a trio of news vans parked in front of the Starbucks, antenna masts projecting from their roofs, a cameraman stared quizzically up at the canyon. Next to the SuperCuts, security guards stood outside two nondescript storefronts; stenciled on the windows were the words "Community Resource Center" and, in smaller letters, "SoCalGas." The guards asked for identification and dismissed anyone who tried to take a photograph. At the entrance to Bath & Body Works, a device that resembled an electronic parking meter was balanced on a tripod; the digital display read "BENZENE," followed by a series of indecipherable ideograms. The parking lot held a

Pre-process

Analyze

Clear



Flesch Kincaid Grade Level 6.7

This text is low in narrativity which indicates that it is less story-like and may have less familiar words. Less story-like texts are usually harder to comprehend. It is high in syntactic simplicity which means that it has simple sentence structures. Simple syntax is easier to process. This text is low in both referential and deep cohesion, suggesting that the reader may have to infer the relationships between sentences and ideas. If the reader has insufficient prior knowledge, these gaps can be challenging.

# REFLECTION

WHAT ARE SOME OF THE BENEFITS TO USING A TOOL LIKE COH-METRIX TO ANALYZE TEXT?

HOW MIGHT THE ANALYSIS OF TEXT FEATURES AND COHESION BE VISUALLY DEMONSTRATED?



IN WHAT WAYS MIGHT EDUCATORS APPLY KNOWLEDGE GAINED FROM AUTOMATED TEXT ANALYSIS? (HOW IS IT DIFFERENT FOR TEXT ANALYSIS VS. STUDENT WRITING?)


# MACHINE LEARNING - FOCUS ON TEXT

Luis von Ahn:

## Massive-scale online collaboration

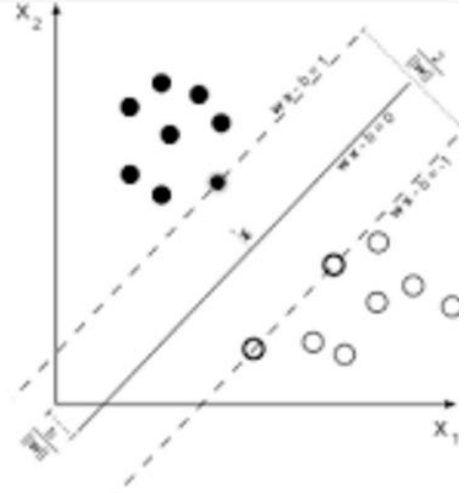
TEDxCMU · 16:39 · Filmed Apr 2011

 33 subtitle languages 

 [View interactive transcript](#)



**Machine learning** is a subfield of computer science that evolved from the study of pattern recognition and computational **learning** theory in artificial intelligence. In 1959, Arthur Samuel defined **machine learning** as a "Field of study that gives computers the ability to learn without being explicitly programmed".



[Machine learning - Wikipedia, the free encyclopedia](https://en.wikipedia.org/wiki/Machine_learning)

[https://en.wikipedia.org/wiki/Machine\\_learning](https://en.wikipedia.org/wiki/Machine_learning) Wikipedia ▾

# LIGHTSIDE TOOLBOX



The open-source LightSide platform, including the machine-learning and feature-extraction core as well as the researcher's workbench UI, has been and continues to be funded in part through Carnegie Mellon University, in particular by grants from the National Science Foundation and the Office of Naval Research. See the full acknowledgements and grant details [below!](#)

We make the LightSide research platform freely available for research and education. In exchange, we ask that you provide us with basic information about who you are and how you're making use of LightSide's capabilities.

Name:

Affiliation:

Email:

# LIGHTSIDE TERMINOLOGY AND KEY FEATURES

**Classifier:** The item category to be tracked. For example, “ is classified as punctuation.

**N-Grams:** The number of times a word appears - unigrams, bigrams, trigrams adjacent to each other.

**Feature Extraction:** The process of finding the features you are searching for and creating a chart.

**Kappa:** A measure between 0 - 1 of accuracy that uses a formula to account for guessing. It is lower than the “accuracy” score.

# ACTIVITY

1. Open LightSide application and select one of the data . csv files that are included in the original download zip folder.
2. Run the analysis of the text.
3. Play with the settings.
4. What different types of data do you get? What does it tell you?

For help: Video overview - <https://www.youtube.com/watch?v=ge3L1DXFHTA>



# BIG QUESTIONS

HOW COULD MACHINE LEARNING TOOLS AND ANALYTICS BE IMPLEMENTED WITH TEXT TO IMPROVE LEARNING?

IS THERE A WAY TO CREATE A “RECAPTCHA” EXPERIENCE?