

What Can We Know About an Anonymous Writer?

Moshe Koppel

with J. Schler, S. Argamon and many students

Yesterday we had our second jazz competition. Thank God we weren't competing. We were sooo bad. Like, I was so ashamed, I didn't even want to talk to anyone after. I felt so rotten, and I wanted to cry, but...it's ok.

Teen Female American

Twenties

Thirties

Male

Yesterday we had our second jazz competition. Thank God we weren't competing. We were sooo bad. Like, I was so ashamed, I didn't even want to talk to anyone after. I felt so rotten, and I wanted to cry, but...it's ok.

Which is Male/Female?

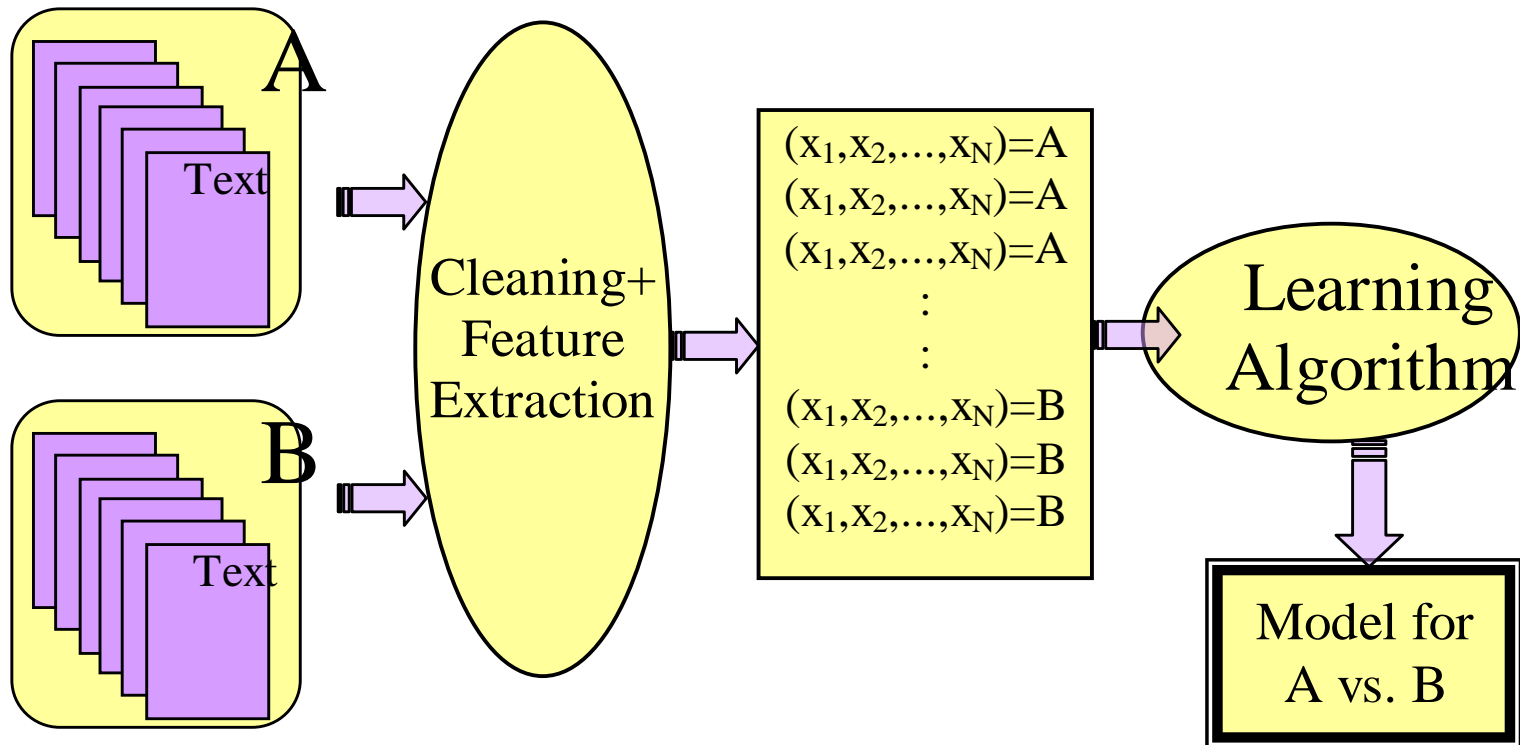
My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .

The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.

Profiling

- We can build classifiers that determine an anonymous writer's:
 - Gender
 - Age
 - Native language
 - Political preferences
 - Ideological/organizational affiliation
 - Personality type
- This is surprisingly accurate and lots of fun (and sometimes embarrassingly stereotypical).
- In fact, we can often determine an anonymous writer's actual identity.

Text Categorization



Features for Categorizing by Style

- All words (BOW)
- Function words (FW)
- Parts of speech (POS)
- Syntax
- Morphology
- Complexity measures
- Idiosyncrasies

In most cases, FW is enough (esp. if we want to ignore content) and BOW is even better (if we want to include content).

Gender (KAS03)

- 566 documents from British National Corpus
- Evenly split among male/female authors
- Controlled for genre and sub-genre

Gender

- 566 documents from British National Corpus
- Evenly split among male/female authors
- Controlled for genre and sub-genre

- Represent documents in terms of FW + POS
- Train using SVM
- Evaluate using 10-fold cross-validation

Gender

- 566 documents from British National Corpus
- Evenly split among male/female authors
- Controlled for genre and sub-genre

- Represent documents in terms of FW + POS
- Train using SVM
- Evaluate using 10-fold cross-validation

- Accuracy: 80+%
- Need separate classifiers for fiction and non-fiction

Summary: Male vs. Female Style

Males use more:

- Determiners (*the*)
- Adjectives
- *of* modifiers (e.g. *pot of gold*)

Females use more:

- Pronouns
- *for* and *with*
- Negation
- Present tense

Which is Male/Female?

My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .

The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.

Which is Male/Female?

My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .

The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.

Which is Male/Female?

My aim in this article is to **show** that given a relevance theoretic approach to utterance interpretation, it is possible to **develop** a better understanding of what some of these so-called apposition markers **indicate**. It will be argued that the decision to **put** something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he **describes** loose apposition as a rhetorical device. However, he does not **justify** this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he **specify** what kind of effects might be achieved by a reformulation or explain how it **achieves** those effects. In this paper I **follow** Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are **made** in unplanned discourse are also **made** in the pursuit of optimal relevance. However, these are **made** because the speaker **recognises** that the original formulation did not **achieve** optimal relevance .

The main aim of this article is to **propose** an exercise in stylistic analysis which can be employed in the teaching of English language. It **details** the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.

Which is Male/Female?

My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does **not** justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. **Nor** does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did **not** achieve optimal relevance .

The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are **not** as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.

+ NEGATION -

Which is Male/Female?

My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .

The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.

Age (SKAP06)

- Corpus of 20,000 blogs
- Labelled for age
- Controlled for gender

- Represent documents by bag-of-words (BOW)
- Wish to determine age interval: *teens/20s/older*

Age

- Corpus of 20,000 blogs
- Labelled for age
- Controlled for gender

- Represent documents by bag-of-words (BOW)
- Wish to determine age interval: *teens/20s/older*

- CV accuracy: **77%**

The lifecycle of the common blogger...

Word	10s	20s	30s
maths	105	3	2
homework	137	18	15
bored	384	111	47
sis	74	26	10
boring	369	102	63
awesome	292	128	57
mum	125	41	23
crappy	46	28	11
mad	216	80	53
dumb	89	45	22

Word	10s	20s	30s
semester	22	44	18
apartment	18	123	55
drunk	77	88	41
beer	32	115	70
student	65	98	61
album	64	84	56
college	151	192	131
someday	35	40	28
dating	31	52	37
bar	45	153	111

Word	10s	20s	30s
marriage	27	83	141
development	16	50	82
campaign	14	38	70
tax	14	38	72
local	38	118	185
democratic	13	29	59
son	51	92	237
systems	12	36	55
provide	15	54	69
workers	10	35	46

Native Language (KSZ05)

- Int'l Corpus of Learner English
- 1290 English essays, evenly split among speakers of 5 languages
 - Bulgarian, Czech, Russian, French, Spanish

Try it yourself. These were written by Russian, French and Spanish speakers, respectively. Can you tell which is which?

In the second part of this author's novel, called *Time Passes*, time has passed indeed and Mrs Ramsay has died.

There are prejudices of small groups, such as homosexuals, immigrants, AIDS diseases, etc. But "political correctness" has had positive and negative consequences.

There is one more kind of film irritating many television viewers - "soap" serials. «*Santa Barbara*» has even won "Oscar" prize.

Native Language

- Int'l Corpus of Learner English
- 1290 English essays, evenly split among speakers of 5 languages
 - Bulgarian, Czech, Russian, French, Spanish
- Features: FW + character n-grams + errors
- Learner: SVM
- 10-fold CV

Native Language

- Int'l Corpus of Learner English
- 1290 English essays, evenly split among speakers of 5 languages
 - Bulgarian, Czech, Russian, French, Spanish
- Features: FW + character n-grams + errors
- Learner: SVM
- 10-fold CV
- Accuracy: **80%**

Some giveaways:

- French speakers switch au-ou (*ou*thor)
- Russian speakers use fewer *the*, switch c-k (*Oskar*)
- Spanish speakers switch q-c (*cu*ality), m-n (*con*fortable)

Let's look back at our examples. Now it's pretty obvious.

In the second part of this **author's** novel, called Time Passes, time has passed indeed and Mrs Ramsay has died.

There are prejudices of small groups, such as homosexuals, **immigrants**, AIDS diseases, etc. But "political correctness" has had positive and negative **consequences**.

There is one more kind of films irritating many television viewers - "soap" serials. «Santa Barbara» has even won **[the] "Oscar"** prize.

Political Preference (DKGU15)

- Given (mostly apolitical) FB posts, determine political preferences
- 450 Hebrew FB pages, evenly split: *left, center, right*
- Features: bag-of-words
- Learner: Bayesian Regression

Political Preference

- Given (mostly apolitical) FB posts, determine political preferences
- 450 Hebrew FB pages, evenly split: left, center, right
- Features: bag-of-words
- Learner: Bayesian Regression
- 10-fold CV accuracy: **91%** (left/right), **79%** (left/center/right)
- Train on party pages, test on individual pages : **82%** (right/left)
- The distinguishing words are exactly the ones you think they are...
- Can probably do a lot better using link analysis

Can We Determine a Writer's Identity? (KSAW12)

- Suppose we have a single document (500 words) and 10,000 suspects – and we aren't even sure if any of them is the real author.
- As a first try, let's assign an anonymous document to whichever author's known writing is most similar (using the usual vector space model).
- We did this to assign blog posts to one of 10,000 bloggers.

Can We Identify an Individual Writer?

- Suppose we have a single document (500 words) and 10,000 suspects – and we aren't even sure if any of them is the real author.
- As a first try, let's assign an anonymous document to whichever author's known writing is most similar (using the usual vector space model).
- We did this to assign blog posts to one of 10,000 bloggers.
- 46% actually were written by the most similar blogger.
- Surprisingly not bad, but we'd hang the wrong guy 54% of the time.
- Can we figure out which attributions are reliable and say “don't know” for the rest?

When does most similar = actual author?

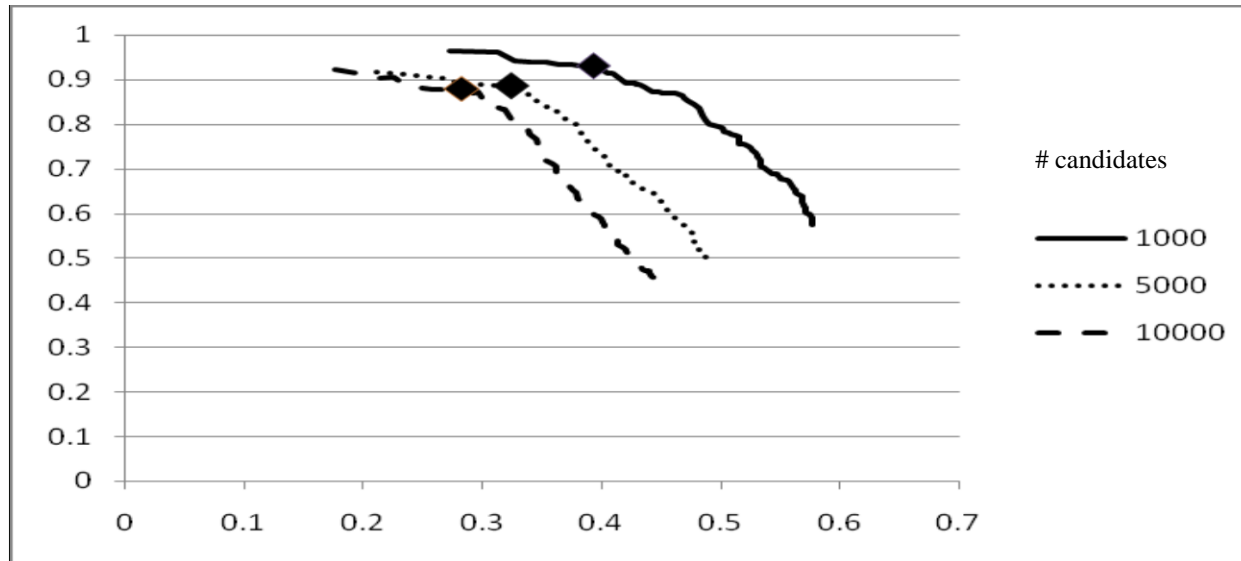
- Check if similarity between post and an author's known text is robust wrt changes in feature set.
 - If it is, that's the author.
 - If not, we just say we don't know.

Formally:

1. Randomly choose subset of features.
2. Find most similar author (using that feature set).
3. Iterate.
4. If S is most similar, for at least k% of iterations, S is author. Else, say Don't Know. (Choice of k trades off precision against recall.)

Results

- 100 iterations, 50% of features per iteration
- training text= 2000 words, post= 500 words



10,000 candidates: 90% precision at 30% coverage ($k=90$)

Who is Satoshi Nakamoto?

- The Holy Grail of attribution:

Identify an author from among millions of candidates (e.g., the Internet)

- The trick: extend the above method using locality sensitive hashing
- Working on it....