Stylometry in plagiarism detection and author profiling

Paolo Rosso
PRHLT Research Center
Universitat Politècnica de València
http://www.dsic.upv.es/~prosso/

Tehran, 25/01/2017
Outline

- Plagiarism
- Intrinsic plagiarism detection
- Author profiling
Plagiarism

- Verbatim
- Paraphrasing
- Ideas
- Cross-language
- Source code
Plagiarism detection

- External: external evidence
- Intrinsic: intrinsic evidence (style analysis)
- Cross-language: translated plagiarism
Intrinsic plagiarism detection

- Insertion of text from a different author into a document causes **style and complexity irregularities**
Stylometry: Intrinsic plagiarism detection

• The study of linguistic style applied to written language
• Quantifying writing style irregularities:

  Text readability: Gunning fog, Flesch–Kincaid, ...
  Vocabulary richness: types/tokens ratio
  Basic statistics: avg. sentence length, avg. word length, word avg. word classes
  n-grams profiles statistics: character level statistics
Gunning fog index

\[ IG = 0.4 \left( \frac{|\text{words}|}{|\text{sentences}|} + 100 \times \left( \frac{|\text{complex_words}|}{|\text{words}|} \right) \right) \]

Complex words: words with three or more syllables

\[ IG(\text{comics}) = 6 \]
\[ IG(\text{Newsweek}) = 10 \]
An example

In this work, we have carried out some research on the influence that mineral salts on the mood of people. For this research I have worked with 5 people who have taken water with different amount of mineral salts. Our theory is that the more minerals are in the water, the more moody people are. [...]

Mineral salts are inorganic molecules of easy ionization in presence of water in living beings they appear by precipitation as well as dissolved mineral salts. [...] Dissolved mineral salts are always ionized. These salts have structural function and pH regulating functions, of the osmotic pressure and of biochemical reactions, in which specific ions are involved.

It seems to me that the results are good. [...]
An example

In this work, we have carried out some research on the influence that mineral salts on the mood of people. For this research I have worked with 5 people who have taken water with different amount of mineral salts. Our theory is that the more minerals are in the water, the more moody people are. [...]

Mineral salts are inorganic molecules of easy ionization in presence of water in living beings they appear by precipitation as well as dissolved mineral salts. [...] Dissolved mineral salts are always ionized. These salts have structural function and pH regulating functions, of the osmotic pressure and of biochemical reactions, in which specific ions are involved.

It seems to me that the results are good. [...]

An example

<table>
<thead>
<tr>
<th>Measure</th>
<th>Global</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>tokens</td>
<td>135</td>
<td>63</td>
<td>72</td>
</tr>
<tr>
<td>types</td>
<td>78</td>
<td>44</td>
<td>46</td>
</tr>
<tr>
<td>W. avg. freq. class</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>avg. sentence length</td>
<td>19.28</td>
<td>21.00</td>
<td>18.0</td>
</tr>
<tr>
<td>avg. word length</td>
<td>4.93</td>
<td>5.38</td>
<td>4.54</td>
</tr>
<tr>
<td>Complexity measures</td>
<td>16.72</td>
<td>17.07</td>
<td>13.82</td>
</tr>
</tbody>
</table>
Intrinsic plagiarism detection @ PAN

- char n-grams (Stamatatos)
- word freq. class + text frequencies (Zechner et al.)
  (Mahgoub et al. @ AraPlagDet)
- Kolmogorov complexity measure (Seaward & Matwin)

...
Intrinsic plagiarism detection @ PAN

- char n-grams (Stamatatos)
- word freq. class + text frequencies (Zechner et al.) (Mahgoub et al. @ AraPlagDet)
- Kolmogorov complexity measure measure (Seaward & Matwin)
  ...
  char n-gram classes based on frequency of n-grams (Bensaleme et al., EMNLP 2015)
Outline

- Plagiarism
- Intrinsic plagiarism detection
- Author profiling
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.

[examples: Moshe Koppel]
British National Corpus

- 920 documents labelled for
  - author gender
  - document genre

- Used 566 controlled for genre

<table>
<thead>
<tr>
<th>Genre</th>
<th>Male</th>
<th>Fem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fiction (prose)</td>
<td>132</td>
<td>132</td>
</tr>
<tr>
<td>Non-fiction</td>
<td>151</td>
<td>151</td>
</tr>
<tr>
<td>Arts (general)</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Arts (acad.)</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Belief/Thought</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Biography</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>Commerce</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Leisure</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Science (gen.)</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Soc. Sci. (gen.)</td>
<td>26</td>
<td>26</td>
</tr>
<tr>
<td>Soc. Sci. (acad.)</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>World Affairs</td>
<td>21</td>
<td>21</td>
</tr>
</tbody>
</table>

Distinguishing features: male vs. female style

Males use more
- Determiners
- Adjectives
- *of* modifiers (e.g. *pot of gold*)

Females use more
- Pronouns *
- *for and with*
- Negation
- Present tense

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 reconstructions are then compared with the original Hemingway version.
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.
Gender: which is female/male?

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 reconstructions are then compared with the original Hemingway version.
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 reconstructions are then compared with the original Hemingway version.
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 reconstructions are then compared with the original Hemingway version.
## Gender & age identification

<table>
<thead>
<tr>
<th>AUTHOR</th>
<th>COLLECTION</th>
<th>FEATURES</th>
<th>RESULTS</th>
<th>OTHER CHARACTERISTICS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argamon et al., 2002</td>
<td>British National Corpus</td>
<td>Part-of-speech</td>
<td>Gender: 80% accuracy</td>
<td></td>
</tr>
<tr>
<td>Koppel et al., 2003</td>
<td>Blogs</td>
<td>Lexical and syntactic features</td>
<td>Gender: 80% accuracy</td>
<td>Self-labeling</td>
</tr>
<tr>
<td>Schler et al., 2006</td>
<td>Blogs</td>
<td>Stylistic features + content words with the highest information gain</td>
<td>Gender: 80% accuracy Age: 75% accuracy</td>
<td></td>
</tr>
<tr>
<td>Goswami et al., 2009</td>
<td>Blogs</td>
<td>Slang + sentence length</td>
<td>Gender: 89.18 accuracy Age: 80.32 accuracy</td>
<td></td>
</tr>
<tr>
<td>Zhang &amp; Zhang, 2010</td>
<td>Segments of blog</td>
<td>Words, punctuation, average words/sentence length, POS, word factor analysis</td>
<td>Gender: 72.10 accuracy</td>
<td></td>
</tr>
<tr>
<td>Nguyen et al., 2011</td>
<td>Blogs &amp; Twitter</td>
<td>Unigrams, POS, LIWC</td>
<td>Correlation: 0.74 Mean absolute error: 4.1 - 6.8 years</td>
<td>Manual labeling Age as continuous variable</td>
</tr>
<tr>
<td>Peersman et al., 2011</td>
<td>Netlog</td>
<td>Unigrams, bigrams, trigrams and tetagrams</td>
<td>Gender+Age: 88.8 accuracy</td>
<td>Self-labeling, min 16 plus 16,18,25</td>
</tr>
</tbody>
</table>
Teams submitting results: 21 (Registered teams: 64)

(Towards) big data: 400,000 social media texts including chat lines of potential pedophiles (task in 2012)

Age classes: 10s (13-17), 20s (23-27), 30s (33-48)

Languages: English and Spanish

http://pan.webis.de/
Approaches: Features

- **Stylistic features**: frequency of punctuation marks, capital letters,...
- Part of Speech
- Readability measures
- Dictionary-based words, topic-based words
- Collocations
- Character or word n-grams
- Slang words, character flooding
- Emoticons
- Emotion words

Author Profiling @ PAN-14: Features

• Similar features of AP@PAN-13:
  content (bag of words, word n-grams) and **stylistic features**

• frequency of words related to different psycholinguistic concepts, extracted from: LIWC and MRC psycholinguistic database

Stylometry: Author profiling

- **Term frequency (F)**: terms with character flooding; terms starting with capital letter; terms in capital letters...
- **Punctuation marks (P)**: frequency of use of dots, commas, colon, semicolon, exclamations and question marks
- **Part-Of-Speech**: frequency of use of each grammatical category
- **Emoticons (E)**: number of different types of emoticons representing emotions
- **Spanish Emotion Lexicon (SEL)**: terms co-occurring with the six basic Ekman’s emotions: happiness, anger, fear, sadness, disgust, surprise
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
I have been taking online courses about valuable subjects that I enjoy studying and that might help me to speak in public.
EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
EmoGraph

He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
I have been taking online courses about valuable subjects that I enjoy studying and that might help me to speak in public.
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
He estado tomando cursos en línea sobre temas valiosos que disfruto estudiando y que podrían ayudarme a hablar en público.

(I) have been taking online courses about valuable subjects that (I) enjoy studying and that might help me to speak in public.
EmoGraph: author’s sentences
## Style +SEL (S) vs EmoGraph (EG)

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Team</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Rangel-EG</td>
<td>0.6624</td>
</tr>
<tr>
<td>2</td>
<td>Pastor</td>
<td>0.6558</td>
</tr>
<tr>
<td>3</td>
<td>Santosh</td>
<td>0.6430</td>
</tr>
<tr>
<td>4</td>
<td>Rangel-S</td>
<td>0.6350</td>
</tr>
<tr>
<td>5</td>
<td>Haro</td>
<td>0.6219</td>
</tr>
<tr>
<td>6</td>
<td>Flekova</td>
<td>0.5966</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>21</td>
<td>Baseline</td>
<td>0.3333</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>23</td>
<td>Mechtli</td>
<td>0.0512</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Team</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Santosh</td>
<td>0.6473</td>
</tr>
<tr>
<td>2</td>
<td>Rangel-EG</td>
<td>0.6365</td>
</tr>
<tr>
<td>3</td>
<td>Pastor</td>
<td>0.6299</td>
</tr>
<tr>
<td>4</td>
<td>Haro</td>
<td>0.6165</td>
</tr>
<tr>
<td>5</td>
<td>Ladra</td>
<td>0.6138</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>8</td>
<td>Rangel-S</td>
<td>0.5713</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>18</td>
<td>Baseline</td>
<td>0.5000</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>23</td>
<td>Gillam</td>
<td>0.4784</td>
</tr>
</tbody>
</table>

Levin’s verb classes

Manual labelling of verbs in Spanish (158) and in English (172) by computational linguists (Autoritas)

- **Emotion**: sentir, querer, amar...
- **Language**: decir, declarar, hablar...
- **Understanding**: entender, saber, conocer, pensar...
- **Perception**: oler, ver, escuchar...
- **Will**: deber, prohibir, permitir...
- **Doubt**: dudar, ignorar...
Levin’s verbs per gender & age

Females vs. Males

To sum up on stylometry

- **Plagiarism detection**: when due to high paraphrasing it is difficult to provide an external evidence of plagiarism, then studying *changes in writing style* could be the only option.

- **Analysis of writing style** could be useful also for tasks such as *author profiling*.
Thanks

Paolo Rosso

prosso@dsic.upv.es