More linguistics for author profiling on Dutch text

Ben Verhoeven
In close collaboration with Walter Daelemans

Presented at Werkgroep Over Taal, VUB
28 April 2016
Introduction

Stylometry

The quantitative study of stylistic characteristics of a text

Writing style

A combination of invariant and unconscious decisions in language production on all linguistic levels, uniquely associated with specific authors or groups of authors

→ Human Stylome Hypothesis (Van Halteren et al. 2005)
Introduction

Computational stylometry

• Authorship identification
  – Attribution - attribute text to one of limited set of authors
  – Verification - is unknown text written by given author?

• Author profiling
  – Prediction of sociological or psychological characteristics of an author
Introduction

Text categorization

• Class representation
• Document representation (features)
• Supervised machine learning method
Class representation

Author profiling
• Age
• Gender
• Location
• Personality
• Education
• Ideology
• Mental health
Document representation

Numeric
• Complexity, readability
• Vocabulary richness
  – Type-token ratio
  – Hapax legomena
• Averages or distributions of
  – Syllable length
  – Word length
  – Sentence length

Character-level
• Letter frequency
• Punctuation
• Spelling errors
• Character n-grams
Document representation

Word-level
• Word n-grams
• Special dictionaries
• Morphology: prefixes and suffixes

Syntax
• Part-of-speech distributions
• Frequencies of syntactic chunks
  (e.g. NP = Det + Adj + N)

...
Which documents?

Data with associated classes needed to train a classifier.
Not that many existing resources (especially for Dutch)

Issues

• Authorial profile can be hard to get
• Not all freely available
  – Non-disclosure agreements
  – Anonymization problems
• None have more than 2 kinds of meta-data
Why do we want all meta-data?

• All aspects have an influence on the author’s writing style

• More importantly: these aspects are reflected in the same kind of features
  – E.g. pronouns (Pennebaker, 2011)

• Solutions:
  – control for some aspects
  – balance the data
  – take all aspects into account
Some resources for personality

- Essays dataset (Pennebaker, later Mairesse)
  - English stream-of-consciousness texts by students
- myPersonality (Stillwell & Kosinski)
  - Large-scale data collection through Facebook app, many languages
- Personae (Luyckx & Daelemans)
  - Dutch essays, written by students
- **CSI Corpus** (Verhoeven & Daelemans)
  - Dutch papers, essays and reviews written by students
- **TwiSty Corpus** (Verhoeven, Daelemans & Plank)
  - Multilingual Twitter stylometry corpus
CLiPS Stylometry Investigation (CSI)

- Corpus in two genres: essays and reviews
- Large amount of meta-data
- Multitude of purposes
  - Mostly in computational stylometry
- Freely available
- Yearly expansion
CSI Corpus

Author meta-data

• Age
• Gender: male/female
• Sexual orientation*: straight or LGBT
• Region of origin: Belgian provinces or The Netherlands
• Personality profile: Big Five and MBTI*

* Provided optionally
Personality typologies

**Big Five**
- Openness to experience
- Conscientiousness
- Extraversion
- Agreeableness
- Neuroticity

Score 0-100 per trait

**MBTI** (Myers-Briggs Type Indicator)
- Extravert – Introvert
- Thinking – Feeling
- Sensing – iNtuition
- Judging – Perceiving

Dichotomy with score 0-100
CSI Corpus

Document meta-data

• Genre
  – Essays, papers: written for Dutch proficiency course (university level), formal text
  – Reviews: special assignment

• Document info
  – Topic, sentiment, veracity (true/false) of reviews
  – Grades of papers and essays
## CSI Corpus

### Corpus size

<table>
<thead>
<tr>
<th>Genres</th>
<th># docs</th>
<th># tokens</th>
<th>Avg. length</th>
<th>Std. dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reviews</td>
<td>1298</td>
<td>202,827</td>
<td>156</td>
<td>65</td>
</tr>
<tr>
<td>Essays</td>
<td>517</td>
<td>565,885</td>
<td>1095</td>
<td>734</td>
</tr>
<tr>
<td>Total</td>
<td>1815</td>
<td>768,712</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CSI Corpus

Advantages
- Multiple purposes
- Yearly expansion
- Text from similar sources (within each genre)
- Enables cross-genre experiments

Disadvantages
- Opportunistic nature (restricted to authors at hand) influences balance of meta-data
Twitter Stylometry (TwiSty)

TwiSty Corpus

– Large-scale multilingual Twitter corpus for personality and gender
– All Western European languages in top 20 of Twitter frequencies, apart from English
  • IT, NL, DE, ES, PT, FR
TwiSty Corpus

• Developed on idea of Plank & Hovy (2015)
  – Twitter mining for only one week
  – Search for MBTI types via API
  – Only English
  – Annotating gender
  – Result
    • 1500 authors
    • 1.2M tweets
Refresher: MBTI

• Myers-Briggs Type Indicator
  – Extraversion vs. Introversion
  – iNtuitive vs. Sensing
  – Thinking vs. Feeling
  – Judging vs. Perceiving

• 16 Types
  – E.g. ESTJ, ISFP, ENTP, ...
TwiSty Corpus

Data collection

• Twitter search instead of mining through API
• Search for combination of each MBTI type with language-specific words
• Download HTML

<table>
<thead>
<tr>
<th>Language</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italian</td>
<td>che, sono, fatto</td>
</tr>
<tr>
<td>Dutch</td>
<td>ik, jij, het, persoonlijkheid</td>
</tr>
<tr>
<td>German</td>
<td>ich, bist, Persönlichkeit, dass</td>
</tr>
<tr>
<td>French</td>
<td>suis, c’est, personnalité</td>
</tr>
<tr>
<td>Spanish</td>
<td>soy, tengo, personalidad</td>
</tr>
<tr>
<td>Portuguese</td>
<td>sou, personalidade</td>
</tr>
</tbody>
</table>
Twisty Corpus

Data clean-up

• Filter out tweets that were not relevant:
  – Not about author
    • @schrooten ok, ik heb deze test destijds met een uitgebreide vragenlijst op mijn werk gedaan. Meerdere van mijn collega PM-ers zijn ESTJ...
  – Ambiguity of type
    • Volgens mij ben ik zowel INTJ als ESTJ -- het eerste als ik me rot voel, het tweede als het goed gaat. #beetjevreemd
  – In different language
    • Estj seregas muzon4ik? Het. O, nu tad davaj daj timati, etoj dj dljee.;D

• Label for gender
TwiSty Corpus

- Corpus size in profiles

<table>
<thead>
<tr>
<th>DE</th>
<th>IT</th>
<th>NL</th>
<th>FR</th>
<th>PT</th>
<th>ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>411</td>
<td>490</td>
<td>1,000</td>
<td>1,405</td>
<td>4,090</td>
<td>10,772</td>
</tr>
</tbody>
</table>

- Corpus size in tweets

<table>
<thead>
<tr>
<th>Language</th>
<th>Total</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>952,549</td>
<td>2,318</td>
<td>819</td>
<td>2,628</td>
</tr>
<tr>
<td>Italian</td>
<td>932,785</td>
<td>1,904</td>
<td>912</td>
<td>2,146</td>
</tr>
<tr>
<td>Dutch</td>
<td>2,083,484</td>
<td>2,083</td>
<td>963</td>
<td>2,426</td>
</tr>
<tr>
<td>French</td>
<td>2,786,589</td>
<td>1,983</td>
<td>932</td>
<td>2,254</td>
</tr>
<tr>
<td>Portuguese</td>
<td>8,833,132</td>
<td>2,160</td>
<td>878</td>
<td>2,456</td>
</tr>
<tr>
<td>Spanish</td>
<td>18,547,622</td>
<td>1,722</td>
<td>952</td>
<td>1,930</td>
</tr>
</tbody>
</table>
TwiSty Corpus

Language Identification

• Many bilingual/polyglot Twitter users
• Tweet-level identification
• Majority voting approach with three language identifiers

<table>
<thead>
<tr>
<th>Tool</th>
<th>Authors</th>
<th># Langs</th>
</tr>
</thead>
<tbody>
<tr>
<td>langid.py</td>
<td>Lui &amp; Baldwin (2012)</td>
<td>97</td>
</tr>
<tr>
<td>langdetect</td>
<td>Nakatani (2010)</td>
<td>53</td>
</tr>
<tr>
<td>ldig</td>
<td>Nakatani (2012)</td>
<td>17</td>
</tr>
</tbody>
</table>
Twisty Corpus

- Corpus size in tweets

<table>
<thead>
<tr>
<th>Language</th>
<th>Total</th>
<th>Confirmed</th>
<th>% Confirmed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italian</td>
<td>932,785</td>
<td>658,332</td>
<td>70.6</td>
</tr>
<tr>
<td>Dutch</td>
<td>2,083,484</td>
<td>1,541,259</td>
<td>74.0</td>
</tr>
<tr>
<td>German</td>
<td>952,549</td>
<td>713,744</td>
<td>74.9</td>
</tr>
<tr>
<td>Spanish</td>
<td>18,547,622</td>
<td>13,493,445</td>
<td>72.8</td>
</tr>
<tr>
<td>French</td>
<td>2,786,589</td>
<td>1,995,865</td>
<td>71.6</td>
</tr>
<tr>
<td>Portuguese</td>
<td>8,833,132</td>
<td>6,353,763</td>
<td>71.9</td>
</tr>
</tbody>
</table>
Experiment

• Instances: 200 tweets per user
• Preprocessing: normalize urls, hashtags, mentions and tokenize
• Features: character and word n-grams
• Model: LinearSVC
• Evaluation: 10-fold cross-validation
## Gender prediction

<table>
<thead>
<tr>
<th>Language</th>
<th>WRB</th>
<th>MAJ</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>50.28</td>
<td>53.75</td>
<td>77.62</td>
</tr>
<tr>
<td>IT</td>
<td>54.78</td>
<td>65.46</td>
<td>73.29</td>
</tr>
<tr>
<td>NL</td>
<td>50.04</td>
<td>51.41</td>
<td>82.61</td>
</tr>
<tr>
<td>FR</td>
<td>51.84</td>
<td>59.60</td>
<td>83.80</td>
</tr>
<tr>
<td>PT</td>
<td>52.15</td>
<td>60.36</td>
<td>87.55</td>
</tr>
<tr>
<td>ES</td>
<td>51.00</td>
<td>57.06</td>
<td>87.62</td>
</tr>
</tbody>
</table>
## Personality prediction

<table>
<thead>
<tr>
<th>Lang</th>
<th>Trait</th>
<th>WRB</th>
<th>MAJ</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE</td>
<td>I-E</td>
<td>60.22</td>
<td>72.61</td>
<td>72.27</td>
</tr>
<tr>
<td></td>
<td>S-N</td>
<td>71.03</td>
<td>82.43</td>
<td>74.49</td>
</tr>
<tr>
<td></td>
<td>T-F</td>
<td>51.16</td>
<td>57.62</td>
<td>59.03</td>
</tr>
<tr>
<td></td>
<td>J-P</td>
<td>53.68</td>
<td>63.57</td>
<td>61.99</td>
</tr>
<tr>
<td>IT</td>
<td>I-E</td>
<td>65.54</td>
<td>77.88</td>
<td>77.78</td>
</tr>
<tr>
<td></td>
<td>S-N</td>
<td>75.60</td>
<td>85.78</td>
<td>79.21</td>
</tr>
<tr>
<td></td>
<td>T-F</td>
<td>50.31</td>
<td>53.95</td>
<td>52.13</td>
</tr>
<tr>
<td></td>
<td>J-P</td>
<td>50.19</td>
<td>53.05</td>
<td>47.01</td>
</tr>
<tr>
<td>NL</td>
<td>I-E</td>
<td>53.02</td>
<td>62.28</td>
<td>62.90</td>
</tr>
<tr>
<td></td>
<td>S-N</td>
<td>57.66</td>
<td>69.57</td>
<td>70.49</td>
</tr>
<tr>
<td></td>
<td>T-F</td>
<td>51.47</td>
<td>58.59</td>
<td>59.95</td>
</tr>
<tr>
<td></td>
<td>J-P</td>
<td>52.00</td>
<td>60.00</td>
<td>57.99</td>
</tr>
</tbody>
</table>
## Personality prediction

<table>
<thead>
<tr>
<th>Lang</th>
<th>Trait</th>
<th>WRB</th>
<th>MAJ</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>I-E</td>
<td>54.77</td>
<td>65.44</td>
<td>66.49</td>
</tr>
<tr>
<td></td>
<td>S-N</td>
<td>68.00</td>
<td>80.00</td>
<td>78.90</td>
</tr>
<tr>
<td></td>
<td>T-F</td>
<td>50.65</td>
<td>55.68</td>
<td>58.22</td>
</tr>
<tr>
<td></td>
<td>J-P</td>
<td>52.13</td>
<td>60.32</td>
<td>56.79</td>
</tr>
<tr>
<td>PT</td>
<td>I-E</td>
<td>53.36</td>
<td>62.97</td>
<td>66.69</td>
</tr>
<tr>
<td></td>
<td>S-N</td>
<td>65.60</td>
<td>76.08</td>
<td>73.42</td>
</tr>
<tr>
<td></td>
<td>T-F</td>
<td>51.27</td>
<td>57.98</td>
<td>61.62</td>
</tr>
<tr>
<td></td>
<td>J-P</td>
<td>50.87</td>
<td>56.61</td>
<td>56.53</td>
</tr>
<tr>
<td>ES</td>
<td>I-E</td>
<td>50.00</td>
<td>50.49</td>
<td>61.09</td>
</tr>
<tr>
<td></td>
<td>S-N</td>
<td>55.42</td>
<td>66.47</td>
<td>61.54</td>
</tr>
<tr>
<td></td>
<td>T-F</td>
<td>51.63</td>
<td>59.04</td>
<td>59.73</td>
</tr>
<tr>
<td></td>
<td>J-P</td>
<td>51.53</td>
<td>58.75</td>
<td>56.08</td>
</tr>
</tbody>
</table>
Conclusion

• Large-scale, “opportunistic”, multilingual social media corpus

• Gender prediction works very well
• Personality prediction is more difficult, yet possible
More linguistics

• Discourse

• Semantics
Discourse

• What
  – relations between sentences
  – coherent structure
  – situating text in the world

• How
  – discourse relational devices (DRD)
Discourse

Features
• Dictionary with categories for different kinds of discourse structure
• Frequencies of categories are an approximation of their use

Source
• Extracted word lists from Dutch Wiktionary
  – 1,300 adverbs
  – 82 conjunctions

Annotation
• Separate annotation for adverbs and conjunctions
• Intuitive annotation by one annotator
• Items can belong to multiple categories
Discourse

Categories for adverbs

• Based on ANS Dutch grammar (Haeseryn et al., 1997)
  – Place/direction: opzij, nergens
  – Time: pas, wanneer
  – Frequency: soms, doorgaans
  – Grade/intensity: erg, nogal
  – Quantification: bijna, ook
  – Manner: graag, anders
  – Modality: misschien, wellicht
  – Negation: nergens, niet
  – Conjunction: immers, trouwens
  – Preposition: buiten, onderin
  – Question: hoe, wanneer
Discourse

Categories for conjunctions

• Based on Penn Discourse Treebank tagset
  (PDTB Research Group, 2007)

  – TEMPORAL
    • Synchronous: terwijl
    • Asynchronous: alvorens, nadat

  – CONTINGENCY
    • Cause: dankzij, want
    • Condition: aangezien, als
Discourse

Categories for conjunctions
• Based on Penn Discourse Treebank tagset
  (PDTB Research Group, 2007)
  – COMPARISON
    • Contrast: oftewel
    • Concession: ofschoon, wanneer
  – EXPANSION
    • Conjunction: alsook, eveneens
    • Instantiation: zoals
    •Restatement: alsof
    • Alternative: noch, hetzij
    • Exception: uitgezonderd
    • List: en
Experiment

Gender prediction
• Given a text, predict the gender of the author

Corpora
• Blogger corpus: 301,080 instances
• CLiPS Stylometry Investigation (CSI) corpus
  (Verhoeven & Daelemans, 2014)
  – Reviews: 1,298 instances
  – Essays: 517 instances (appeared not to be enough)
Experiment

Association analysis
• Correlation analysis relating numerical to binary variable
  – Predictors (numerical): relative counts of discourse categories
  – Outcome (binary): gender

Logistic regression
• Relative features are normalized
• Fit binomial glm
• Coefficients converted to probabilities
• Computed 95% and 99% two-tailed confidence intervals for statistical significance
Results

Probabilities (of class 1)
  – Reviews: between 0.070 and 0.407
  – Blogs: between 0.229 and 0.337

So all are more related to female (since male = 1)

Interpretability
  • Which gender uses which category more?
  • How strong is the association?
## Results

<table>
<thead>
<tr>
<th></th>
<th>Reviews M</th>
<th>Reviews F</th>
<th>Blogs M</th>
<th>Blogs F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concession</td>
<td>152.04</td>
<td>160.92</td>
<td>61.17</td>
<td>61.98</td>
</tr>
<tr>
<td>Alternative</td>
<td>30.55</td>
<td>31.56</td>
<td>14.39</td>
<td>14.31</td>
</tr>
<tr>
<td>Exception</td>
<td>0.00</td>
<td>0.00</td>
<td>0.0050</td>
<td>0.0035</td>
</tr>
<tr>
<td>Comparison</td>
<td>154.91</td>
<td>162.93</td>
<td>61.41</td>
<td>62.21</td>
</tr>
<tr>
<td>Condition</td>
<td>84.73</td>
<td>72.71 *</td>
<td>24.06</td>
<td>23.54 **</td>
</tr>
<tr>
<td>Expansion</td>
<td>360.65</td>
<td>373.41</td>
<td>149.22</td>
<td>149.33 **</td>
</tr>
<tr>
<td>Instantiation</td>
<td>11.46</td>
<td>10.04 *</td>
<td>2.776</td>
<td>2.712</td>
</tr>
<tr>
<td>Restatement</td>
<td>12.65</td>
<td>13.49 *</td>
<td>3.293</td>
<td>3.218</td>
</tr>
<tr>
<td>Place</td>
<td>745.18</td>
<td>764.88 *</td>
<td>296.49</td>
<td>294.60</td>
</tr>
<tr>
<td>Preposition</td>
<td>802.70</td>
<td>755.41 **</td>
<td>281.24</td>
<td>276.30 **</td>
</tr>
<tr>
<td>Question</td>
<td>48.93</td>
<td>56.84</td>
<td>15.19</td>
<td>14.81</td>
</tr>
<tr>
<td>Manner</td>
<td>576.90</td>
<td>567.70</td>
<td>210.53</td>
<td>211.55 **</td>
</tr>
<tr>
<td>Frequency</td>
<td>25.54</td>
<td>32.87</td>
<td>8.75</td>
<td>8.97</td>
</tr>
<tr>
<td>Negation</td>
<td>103.35</td>
<td>117.07</td>
<td>31.47</td>
<td>31.05 **</td>
</tr>
</tbody>
</table>
Conclusion

- Some significant associations, yet several are weak
- More significance in blogs (bigger corpus)
- Two categories relevant in both corpora:
  - Condition and Preposition are both used more by men
- Frequency differences between corpora related to different genres
Thanks for your attention

Questions?

ben.verhoeven@uantwerpen.be
@verhoevenben