Studies of Disputed Authorship

Michael P. Oakes

Chapter 1 of “Literary Detective Work on the Computer”, John Benjamins.

Authorial “fingerprints”

- Automatic author identification is a branch of computational stylometry, which is the computer analysis of writing style.
- It is based on the idea that an author’s style can be described by a unique set of textual features, typically the frequency of use of individual words, but sometimes considering the use of higher level linguistic features.
- Disputed authorship studies assume that some of these features are outside the author’s conscious control, and thus provide a reliable means of discriminating between individual authors.
- Holmes (1997) refers to this as the “human stylome”.
- There is no definitive proof that such features exist, and thus the field of automatic authorship attribution lacks a strong theoretical foundation.

Authorial “fingerprints” (2)

- If such a distinctive “stylistic signature” or “authorial fingerprint” does exist, it would most likely be made up of many weak discriminators (such as the frequencies of individual words) rather than a few strong hard and fast rules (Burrows, 2002)
- Burrows (1992:91) had earlier warned that “No one has yet identified a stylistic attribute as idiosyncratic or as durable as human fingerprints. Nothing in the nature of the case suggests that anyone will ever do so”.
- but as we shall see, many studies have successfully made use of high frequency function words like “the”, “of” and “and”, which tend to have grammatical functions rather than reveal the topic of the test.
- Their usage is unlikely to be consciously regulated by authors, and tends to be constant with theme.

The closed class situation

- Author identification studies are easier in “closed class” situations, where the text to be attributed could only have been written by one of a small number of plausible candidates.
- These studies require large samples of texts undisputedly written by each of the candidate authors.
- In the case of two candidates, we will call these samples corpora A and B. Corpus C is the disputed text.
- A set of features and a suitable statistical measure is chosen which reliably discriminates between the two candidate authors.
- Using the same set of features and statistical measure set of features, we determine whether corpus C is more similar to corpus A or corpus B.

Authorship attribution “in the wild”

- However, the situations that must be dealt with in practice are often not so straightforward, and Koppel et al. (2011) list three important, but more difficult scenarios:
- There may be thousands of known candidate authors.
- The author of the anonymous text might be none of the known candidates.
- The known text for each candidate and/or the anonymous text might be very limited.

… in the wild (2)

- Many candidate authors – distance measures more appropriate than machine learning.
- Author of unknown texts might be none of the known candidates – allow system to return “don’t know”.
- For a pair of short texts, X and Y, their suggested approach is to choose a set of “impostors” of Y (roughly similar texts chosen by a method such as the search-engine technique above, based on their intertextual distance such as the cosine similarity to Y, see section 3.3). It is a little bit like an identification parade at a police station – if X is compared to the set of Y and all its impostors, and correctly “chooses” Y by being found most similar to it, we can conclude that the two texts are by the same author. Koppel et al. describe this last approach as “somewhat speculative”.

Preprocessing the texts and the vector space model

- Using the methodology of text classification, we start by cleaning the corpus, our electronically-stored set of texts.
- This involves making decisions such as where do the word boundaries occur, and should we reduce all words with initial capitals to lower case? (Juola, 2006).
- The next step is to represent texts as numerical vectors as a search engine might do.

“the”	“of”	“and ”	Noun	Com ma
100	90	80	70	10

Choice of features

- A number of recent authors use long lists of "most frequent words" (MFW) such as Hoover (2007) who considers the top 4000, but Burrows (2002) used only the most frequent 40 to 150.
- The features are chosen due to their ability to help discriminate between authors' writing styles.
- Making use of inter-text distance measures, we are then able to compare the vectors for our unknown text with each of the samples of known authorship, and find the best matching known sample, which was most probably written by the same author.

Text classification techniques

- A group of techniques, placed by Harald Baayen (2008) under the umbrella term of “clustering”, not only calculate distances between texts, but produce maps of where a set of texts lie in relation to each other.
- We will also briefly compare the performance of machine-learning approaches to those based on inter-textual distances, since in the simplest “closed set, adequate amounts of training data” case, successful techniques tend to be either distance measure-based and machine-learning methods.

Support Vector Machines

- SVMs are suitable for classifying texts represented either as binary vectors (where each 1 or 0 represents the presence or absence of a word in a text) or as vectors containing the exact numbers of times each word was found in the texts.
- The SVM is trained by a process called supervised learning, meaning that examples of texts and their true authors are input until the SVM is able to distinguish between vectors typical of different writers.
- Vectors representing unknown texts can then be input, and the SVM will automatically classify them as being more typical of a particular author.

SVM Case study

- An SVM was used to examine the mystery of an unfinished work by the Romanian novelist Mateiu Caragiale.
- After Caragiale's death, another author, Radu Albala, claimed to have found the "lost" conclusion, but later admitted to have written it himself.
- Dinu and Popescu (2009) used an SVM classifier to show that texts by Caragiale and Albala could be distinguished automatically, and that the "lost" conclusion was indeed written by Albala.

Choice of features

- The textual features chosen to distinguish writing styles must both be common enough to demonstrate statistical significance, and objectively measurable or countable.
- The earliest features to be proposed were word and sentence length, as described in a letter dated 1887 by Mendenhall (Kenny, 1982).
- However, these measures are under conscious control of the author, and may be better discriminators of genre or register. For example, word and sentence length will be greater on average in a quality newspaper than in a traditional tabloid.
- Another approach is simply to use a fixed number (such as 100) of the most common words in the combined corpus of all the texts under consideration.

Choice of features (2)

- Other authors such as Kjell (1994) have used substrings of words.
- DeForest and Johnson (2001) used the proportion of English words of Latinate origin to those of Germanic origin to discriminate between the characters in Jane Austen's novels, Latinate words being considered to be more suggestive of high social class, formality, insincerity and euphemism, lack of emotion, maleness and stateliness.
- If syntactically annotated corpora are available, analyses above the lexical level are possible. Antosch (1969) showed that the ratio of adjectives to verbs was higher in folk tales than scientific texts. Baayen et al. (1996) counted the frequency with which each phrase rewrite rule was used in parsing a corpus of crime fiction to distinguish the styles of two writers.

Vocabulary richness

- Very early studies used a single measure, hoped to be invariant within the writings of a single author, such as average word length and statistics based on the type-token ratio.
- The number of word types in a text is the number of unique words, while the number of tokens is the total number of words.
- In the title “Men are from Mars, Women are from Venus”, there are 6 words types but 8 tokens overall (“are” and “from” are repeated). The type-token ratio is 6 / 8 = 0.75.
- In general, the vocabulary is rich if many new words appear in a portion of text of a certain length, but is poor if relatively few distinct words appear in a text of that length.
- Other measures of VR include Yule’s K.

Feature reduction

- One approach to feature selection is to initially consider all possible features of a text, and winnow them down to a smaller set of discriminators all of which work well.
- Yang and Pedersen (1997), writing about text classification in general, describe five statistical measures: document frequency DF, information gain IG, mutual information MI, a chi-squared statistic (CHI) and their own term strength (TS).
- Initially every word in the texts is a potential discriminator and given a score by one of these measures, but only the best scoring words are retained as features.

Pointwise Mutual Information

A = number of documents containing term t found in category c	B = number of documents containing term t not found in category c
C = number of documents which do not contain term t in category c	D = number of documents which do not contain term t nor are in category c

- PMI can then be estimated by
- $I(t,c) = log_2 \frac{A \times N}{(A+C)\times(A+B)}$

t-test for independent samples

- Binongo and Smith (1999) employed another commonly-used statistical test, the t-test for independent samples, to find the best 25 discriminators between ten texts by Shakespeare and five texts by Wilkins.
- Imagine that we put the frequency of the word “then” in each of the Shakespeare texts and each of the Wilkins texts into two lists using the R programming language, as follows:
- shakespeare = c(10, 9, 8, 7, 11, 8, 12, 13, 6, 10)
- wilkins = c(14, 15, 20, 16, 18)
- t.test(shakespeare, wilkins, paired=F)
- This returns a t value of –5.60. Imagine that we repeat the experiment for “by” and obtain a t value of 3.88. Since the absolute value for “then” was greater, the word “then” must be a better discriminator between the writings of Shakespeare and Wilkins than the word “by”.

Jack Grieve's (2007) comparison of feature sets.

Textual Measurement (variant)	40	20	10	5	4	3	2
Word and punctuation mark profile (5-limit)	63	72	80	87	89	92	95
2-gram profile (10-limit)	65	72	79	86	88	91	94
3-gram profile (10-limit)	61	72	78	85	88	91	94
4-gram profile (10-limit)	55	64	73	83	85	89	93
Character and punctuation mark profile (5-limit)	50	60	70	81	84	87	93
Multiposition character profile (first and last six in word) thus encompassing most positions	49	58	68	79	82	86	92
Word profile (5-limit)	48	57	67	77	80	85	88

Inter-textual distances

- Having discussed the choice of linguistic features to characterise each of our texts, we now consider methods for deciding how similar (or different) pairs of texts are based on how much they overlap in the possession of common linguistic features.
- This is particularly useful when we want to know if a text of unknown authorship resembles more the known works of author A or author B.
- The similarity property means that if the two texts are identical, their distance should be 0. The symmetry property means that the distance between text A and text B should be exactly the same as that between text B and text A. Finally, the triangle inequality property means that the total distance from A to C plus that from C to B must be equal or greater to the “direct” distance between A and B.
- Some more sophisticated measures of inter-textual distance have an upper limit of 1 for cases where two texts have absolutely no words in common. This has the advantage of allowing a distance measure to be expressed as a similarity simply by subtracting it from 1.

Manhattan distance and Euclidean distance

- The Manhattan distance (or City block distance) between two texts is found by finding the absolute differences in the frequencies of each word type, then adding all these together.
- The distance gets its name because since we cannot cut across city blocks, the shortest distance between two points in a built-up city would be to walk around them.
- $D(x, y) = \sum_{i=1}^{m} |x_i - y_i|$

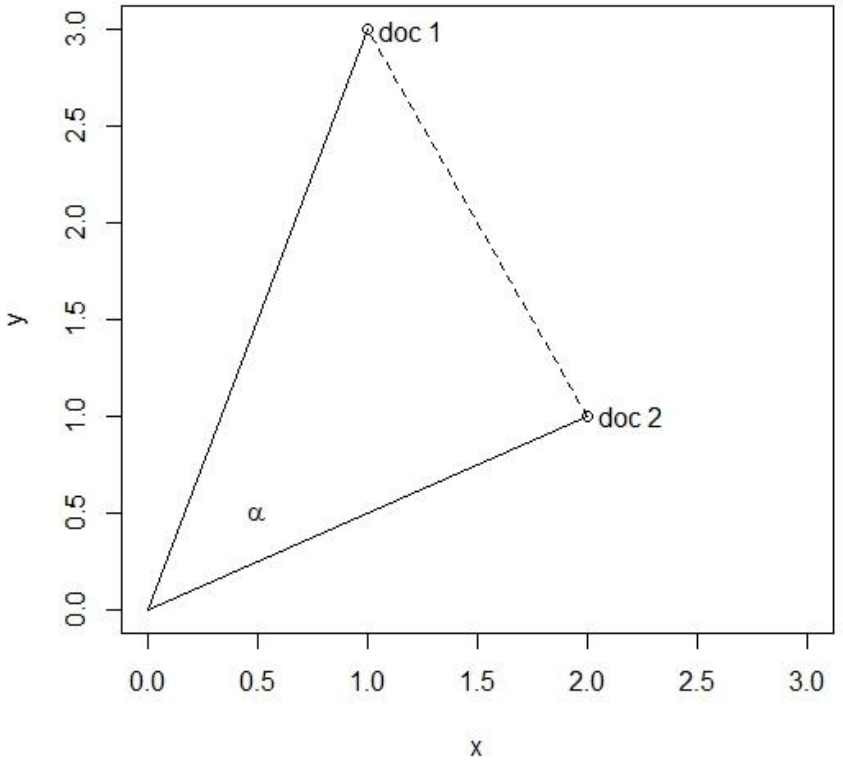
- Euclidean distance, in contrast, corresponds to distance as the crow flies.
- We square the absolute differences in the frequencies of each word type, then add them all together, and finally take the square root of this sum.
- $D(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$
-

Cosine similarity measure

- The similarity between two texts is found by the following formula:

- $$Cosine(doc_1, doc_2) = \frac{\sum_{k=1}^{t}(term_{ik}, term_{jk})}{\sqrt{\sum_{k=1}^{t}(term_{ik})^2 \cdot \sum_{k=1}^{t}(term_{jk})^2}}$$

- where $term_{ik}$ and $term_{jk}$ are the frequencies of word k in documents i and j respectively.



Burrows' Δ : relative frequencies (%)

	Die	Der	Das	Ist	Und	Nicht
BTh Allgemeinheit	2.675	2.551	1.673	1.993	2.107	1.942
BTh Bedeutung	3.284	2.996	2.718	2.123	1.706	1.498
BTh Erwartung	2.852	2.721	2.545	2.583	1.591	1.968
SCH Positivismus	2.608	3.048	1.045	1.607	1.941	1.309
Mean	2.855	2.829	1.995	2.077	1.836	1.679
Standard deviation	0.304	0.234	0.781	0.403	0.232	0.328

$$z = \frac{(NF - mean)}{SD}$$

	Die	Der	Das	Ist	Und	Nicht
BTh Allgemeinheit	-0.592	-1.187	-0.412	-0.207	1.167	0.801
BTh Bedeutung	1.412	0.713	0.925	0.115	-0.561	-0.552
BTh Erwartung	-0.009	-0.459	0.703	1.259	-1.058	0.881
SCH Positivismus	-0.810	0.933	-1.216	-1.166	0.452	-1.130

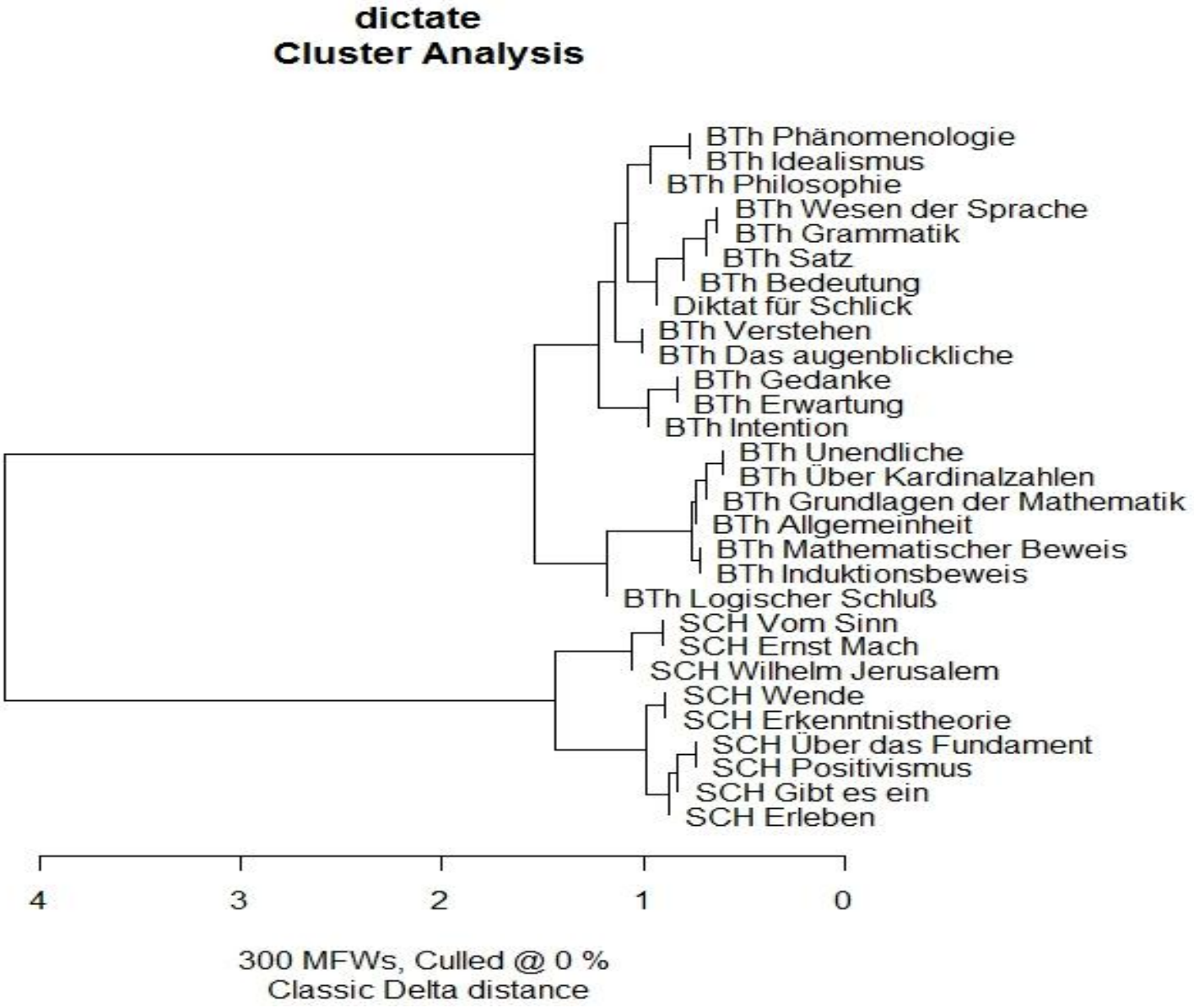
Example Calculation of Burrows' Δ between "BTh Allgemeinheit" and "BTh Bedeutung"

Word	Z(A)	Z(B)	Z(A) – Z(B)	| Z(A) – Z(B) |
Die	-0.592	1.412	- 2.004	2.004
Der	-1.187	0.713	- 1.900	1.900
Das	-0.412	0.925	- 1.337	1.337
Ist	-0.207	0.115	- 0.322	0.322
Und	1.167	-0.561	1.728	1.728
Nicht	0.801	-0.552	1.353	1.353
Total				8.644
Total / N				1.441

Δ scores for each document pair in the corpus

	BTh Allgemeinheit	BTh Bedeutung	BTh Erwartung	SCH Positivismus
BTh Allgemeinheit	0	1.441	1.033	1.125
BTh Bedeutung	1.441	0	0.981	1.243
BTh Erwartung	1.033	0.981	0	1.676
SCH Positivismus	1.125	1.243	1.676	0

Diktat für Schlick



Evaluation of feature-based measures for inter-textual distance (1).

- Forsyth and Sharoff (2013) provide a test bed of 113 documents for the comparison of feature weighting schemes and measures of inter-textual distance. The test bed is available in five languages: German, English, French, Russian and Chinese. The version of this test bed released in April 2012 may be found at http://corpus.leeds.ac.uk/tools/5gcorpus.zip
- To find "gold standard" values for inter-text distances against which each of the "low-level", feature-based techniques could be compared, they did not use any mathematic measure but instead used the answers from human annotators to 17 questions about the nature of the texts.

Evaluation of feature-based measures for inter-textual distance (2).

- Examples of the questions were: To what extent is the text concerned with expressing feelings or emotions? To what extent is the text's content fictional? and To what extent do you judge the text to be modern?
- Responses were given on a 4-point scale, where 0 meant the attribute was absent; 0.5 the attribute was only present to a small extent; 1 the attribute was somewhat or partly present; and 2 the text was strongly characterised by the attribute in question.
- One calculation of the document similarity "gold standard" they used was the inverse product-moment correlation of the mean annotators' responses to two texts.
- The cosine similarity was outperformed by Pearson product-moment correlation, Spearman's rho, and the tetrachoric correlation coefficient (Upton and Cook, 2008).

Tf-idf

- In terms of weighting the features according to their relative importance, transforming raw frequencies to the tf-idf measure widely used in search engine technology worked well. However, good results were also obtained with the simpler technique of binarisation with respect to median frequency. One variant of the formula for tf-idf (Manning et al., 2008:118) is as follows:
- $\text{tf-idf}_{t,d} = tf_{t,d}.log\left(\frac{N}{df_t}\right)$
- Here $tf_{t,d}$ is the frequency of term t in document d, N is the total number of documents in the collection and df_t is the total number of documents in the collection. The idea is that the terms which are most important in document d are those which are frequent in that document, but are not found in many other documents.

Introduction to factor analysis: correlations between linguistic features (Biber).

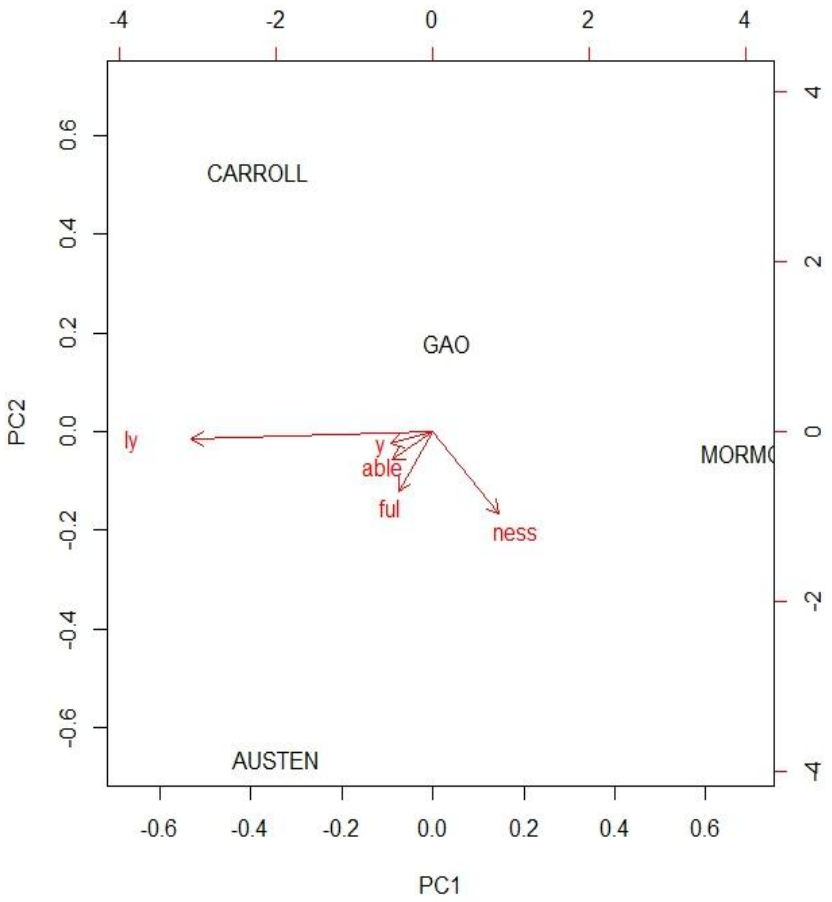
	1st pers. pronoun	questions	passives	nominalisations
1 pers. pronoun	1			
Questions	.85	1		
Passives	-.15	-.21	1	
nominalisations	.08	-.17	.90	1

Factor analysis (2)

- Correlations close to 1 show that two linguistic features tend to vary together – here texts which have large numbers of nominalisations also tend to have large counts of passives.
- Intuitively we can see two underlying “factors” in this data, or groups of features which tend to vary together. One factor is the related pair of questions and first person pronouns, and the other is the pair passives and nominalisations.
- This means we can summarise the original data which considered four different linguistic features (and thus consisted of four dimensions) by describing it by a smaller number of two dimensions or underlying constructs.
- Factor analysis is a computational technique for discovering such underlying factors or “components” automatically from large numbers of initial features.
- The factors are ordered, so that the first one to be extracted explains most of the variation in the original data set.

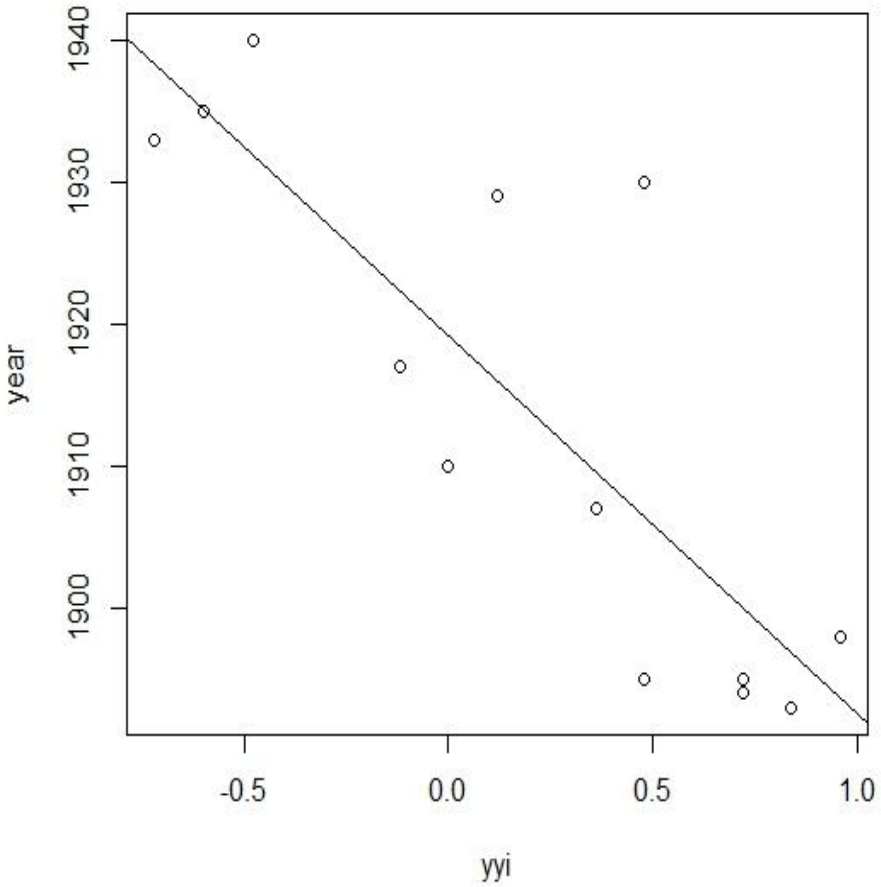
Subset of Baayen's data for a measure of productivity for 5 suffixes in 4 samples of text.

	-ful	-y	-ness	-able	-ly
MORMON	0.188	0.566	2.075	0.000	2.264
AUSTEN	1.289	1.565	1.657	1.012	6.261
CARROLL	0.271	1.087	0.271	0.407	6.385
GAO	0.330	1.983	0.826	0.826	4.462



Stylochronometry (Forsyth 1999):
YYI = (YY – OY) / (YY + OY)

Rank	Substring	Chi-squared	Frequency in poems <=1915	Frequency in poems >= 1916
1	‘what	35.11	30	100
2	‘ can	34.38	21	82
3	‘s, an	25.49	63	19
4	‘ whi	25.44	67	21
5	‘ with	22.30	139	74
6	‘?	21.97	30	83



Conclusions

- Although computational stylometry techniques can be very powerful, findings must always be evaluated in the light of traditional methods of authorship attribution, particularly historical evidence and non-computational studies of style and content (Holmes and Crofts, 2010).
- Holmes and Crofts refer to the 'Burrows' approach as the "first port-of-call for attributional problems". Here the N (typically 50 to 75) most common words in the whole set of documents being compared is taken, and the relative frequency of each of these words in each individual text sample is found. Text samples are typically about 3000 words. This data becomes the input to a multivariate statistical technique, such as cluster analysis or PCA.
- Burrows' Δ is also a currently popular technique.
- Evaluation at PAN-13.