Who Wrote this Novel? Authorship Attribution Across Three Languages

J. Savoy
University of Neuchatel
Computer Science Dept.

Juola P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3).

Love, H. (2002). *Attributing Authorship: An Introduction*, Cambridge University Press, Cambridge, 2002.

Craig H., Kinney A.F.(2009) Shakespeare, Computers, and the Mystery of Authorship, Cambridge, Cambridge University Press.

Authorship Attribution

- Long tradition of research (predating computer science)
- Interest in
  - resolving issues of disputed authorship
  - defining the stylistic elements of a given author
  - identifying authorship of anonymous texts
  - may be useful in detecting plagiarism
  - used in forensic setting (e.g. to detect genuine confessions)
  - other applications related to e-mails, terrorist, …

Authorship Attribution

- One text = one author?
- Collaborative authorship (solitary authorship not often accurate) e.g., Shakespeare's plays
- Precursory authorship (the source or influence)
- Declarative authorship (T. Sorenson behind J.F. Kennedy)

- Not only text! (image, picture, music, …)

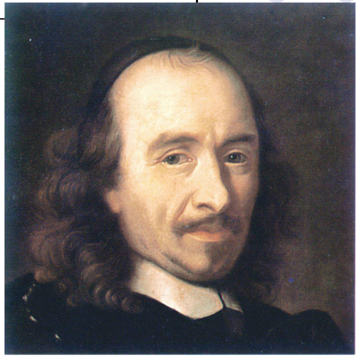
- Focus only on literary works

Some Classical Examples

- Did Shakespeare write all his plays?
  - Various authors including Bacon and Marlowe are said to have written parts or all of several plays
  - “Shakespeare” may even be a nom-de-plume for a group of writers?
- Plays written by more than one author
  - *Edward III* – Shakespeare? & Kyd?
  - *Two Noble Kinsmen* – Shakespeare & Fletcher
  - *Timon of Athens* – Shakespeare & Middleton?
  - *Henry VIII* – Shakespeare & Fletcher?

Craig, H. & Kinney A.F. (Eds): *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge Univ. Press, 2009

Some Classical Examples



- The debate *Molière* vs. *Corneille*?
  Jean Baptiste Poquelin (1622-1673)
  Pierre Corneille (1606-1684)
- *Psyché* (1671), both are authors
- Plays (comedies) from 1658
- Corneille needs money, well-known for his dramas (but cannot write comedies, and inferior genre)
- Pierre Louys (1919) (and Voltaire) indicates that Corneille was the real author based on the rhythmus, versification.

Labbé, D. (2009). Si deux et deux font quatre,
Molière n'a pas écrit Dom Juan. Paris, Max Milo.

Some Modern Examples

- The *Federalist Papers* (Mosteller and Wallace, 1964)
  - A series of articles published in 1787-88 with the aim of promoting the ratification of the new US constitution. Papers written under the pseudonym "Publius"
  - Some are of known (and in some cases joint) authorship but others are disputed
  - Written by three authors, Jay (5), Hamilton (51) and Madison (14), three by Hamilton & Madison, 12 uncertain.
  - Pioneering stylometric methods were famously used by Mosteller and Wallace in the early 1960s to attempt to answer this question
  - It is now considered as settled
  - The *Federalist Papers* present a difficult but solvable test case

How?

- Authorship attribution
  - External evidence (incipits, colophon, biographical evidence, earlier attributions, social world within which the work is created, …)
  - Internal evidence (self-reference, evidence from themes, ideas, beliefs, conceptions of genre, …)
  - Bibliographical evidence
  - Historical, physical evidence
- Stylometry (fingerprint)
  Computer science provides a (quantitative) tool
- "When you can measure what you are speaking about, and express it in numbers, you know something about it"
  Lord Kelvin

Stylometry

- Measurement of (aspects) of style

  "The stylometrist therefore looks for a unit of counting which translates accurately the 'style' of the text, where we may define 'style' as a set of measurable patterns which may be unique to an author?"

  H. Holmes, Authorship Attribution, *Computers & Humanities*, 1994, p. 87
- Assumes that the essence of the *individual style* of an author can be captured with reference to a number of quantitative criteria, called *discriminators*
- Obviously, some aspects of style are conscious and deliberate
  - as such they can be easily imitated and indeed often are
  - many famous pastiches, either humorous or as a sort of homage
- Computational stylometry is focused on *subconscious* elements of style less easy to imitate or falsify

Stylometry

- How?
  - A single measurement
  - Multivariate analysis
  - Text Categorization
    (larger set of the vocabulary)
  - Others (syntax, layout, …)

Single Measurement

- Letter counts
- "What disturb me in Shakespeare's plays is the over-used of the letter "o". I can live with a lot of "e" or "I", but not a lot of "o". So, yes clearly, I prefer reading Marlowe."

Letter Counts

- T. Merriam reports
  "of counting the letters in the 43 plays was the implausible discovery that the letter 'o' differentiates Marlowe and Shakespeare plays to an extent well in excess of chance" (used also letter 'a')
- Frequency less than 0.0078, 6 plays of Marlowe
  Frequency greater than 0.0078, 36 plays of Shakespeare

T. Merriam: Letter Frequency as a Discriminator of Authors. *Notes & Queries*, 239, 1994, p. 467-469.
T. Merriam: Heterogeneous Authorship in Early Shakespeare and the Problem of *Henry V*. *Literary and Linguistic Computing*, 13, 1998, p. 15-28.

Single Measurement

- Letter counts
- Word length
- Sentence length (too obvious and easy to manipulate)
- Frequencies of letter pairs (*n*-gram)
- Distribution of words of a given length (in syllables), especially *relative frequencies*
- Simple, but really effective?

Multivariate Analysis

- Thanks to computers it is now possible to collect large numbers of different measurements, of a variety of features
- Variants of multivariate analysis
  - Principal components analysis (PCA)
  - Correspondence analysis (CA)
  - Cluster analysis
  - …
- Variables = features = word types or lemmas
- Objects = text excerpts

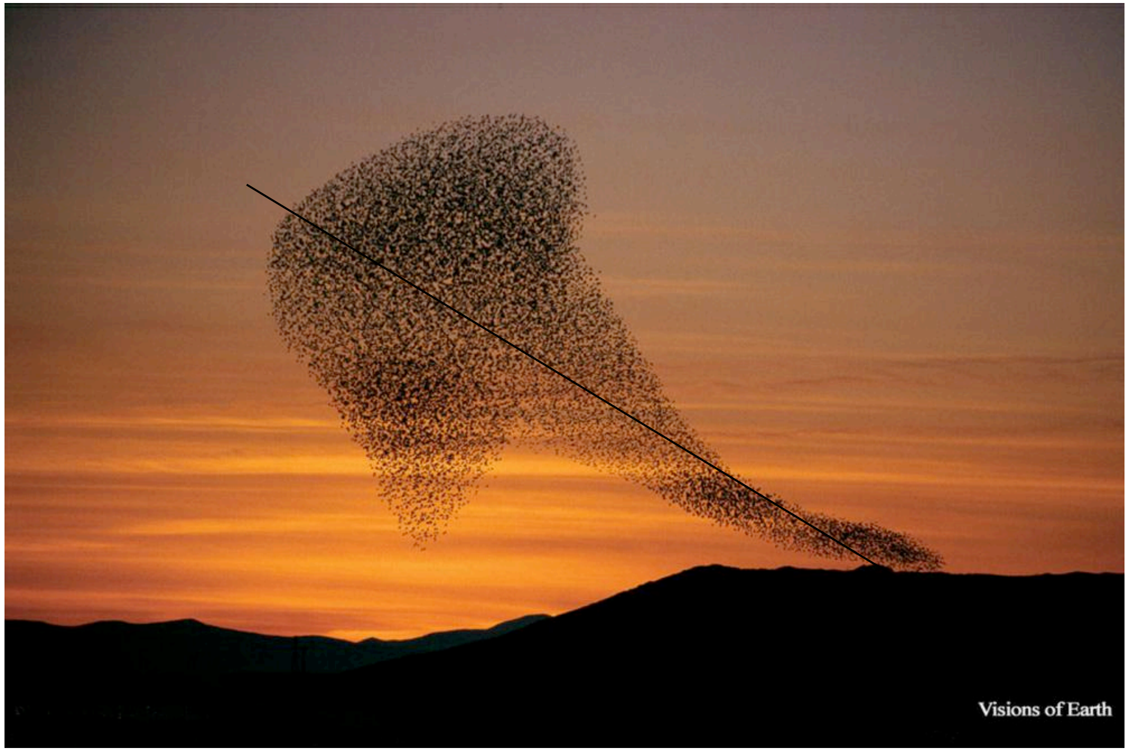
Lexical Table (Small Example)

Occurrence frequency of the most frequent German lemmas

	G1	G3	N25	N27	M39	M40	K42	K43	New
d	665	775	573	894	681	836	758	775	1162
.	345	254	267	318	348	398	351	363	362
und	258	307	323	148	443	473	197	201	183
sein	219	276	258	262	327	262	270	288	178
ich	172	426	203	309	98	48	220	151	1
in	122	133	63	182	177	183	95	124	296
nicht	105	97	128	107	81	52	152	130	66
werden	74	54	35	81	39	44	85	66	85

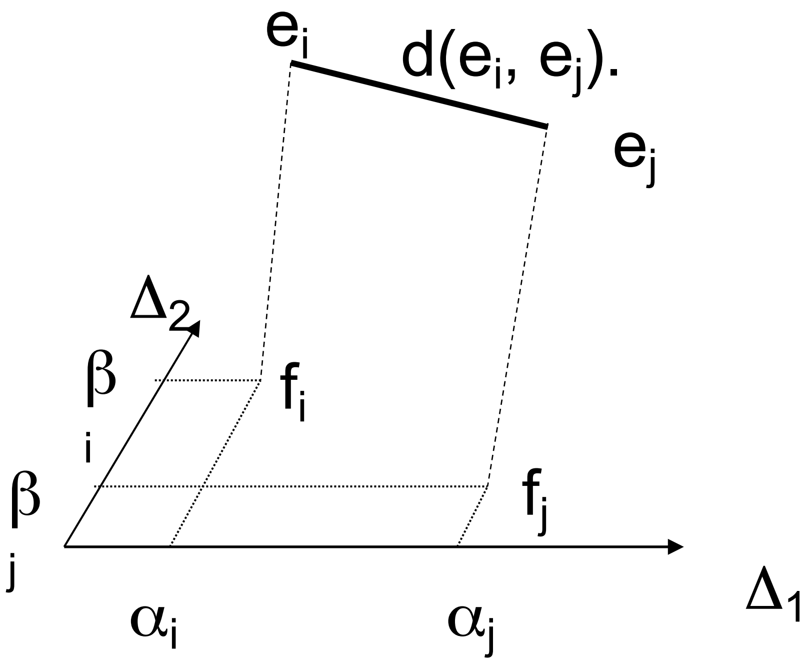
Other Representation

A cloud of birds in 3D → 2D (→ 1D)



Principal Component Analysis

- PCA is a statistical method for arranging large arrays of data into interpretable patterning match
- “principal components” are computed by calculating the *correlations* between all the variables, then grouping them into sets that show the most correspondence



We will define a projection plane (defined by the lines Δ_1 and Δ_2, *perpendicular* (no correlation)) to represent the objects (e_i, e_j) and conserving the real distance $d(e_i, e_j)$.

Lexical Table (Small Example)

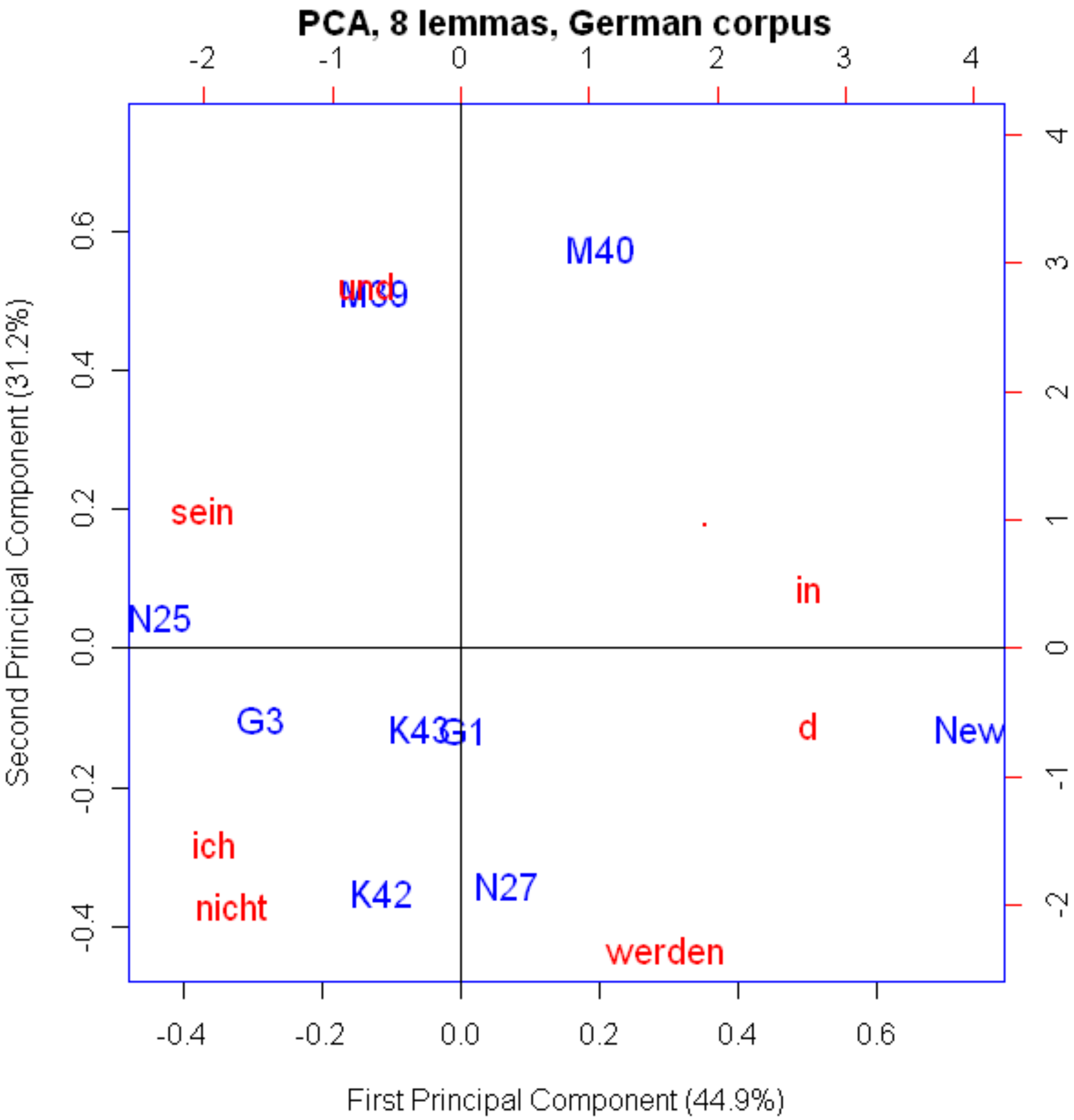
To represent this information into 2D!

	G1	G3	N25	N27	M39	M40	K42	K43	New
d	665	775	573	894	681	836	758	775	**1162**
.	345	254	267	318	348	398	351	363	362
und	258	307	323	148	**443**	**473**	197	201	183
sein	219	276	258	262	327	262	270	288	178
ich	172	**426**	203	309	98	48	220	151	**1**
in	122	133	63	182	177	183	95	124	**296**
nicht	105	97	**128**	107	81	52	**152**	**130**	66
werden	74	54	35	**81**	39	44	**85**	**66**	85

PCA

8 lemmas
(German)

und (T. Mann),
nicht, *werden*
(Kafka)



Corpora

- Three languages
  - German
  - English
  - French
- Literary works (novels, mainly 19th century)
  - Extracted from the Gutenberg Web site
  - Text excerpts of around 10,000 word tokens
- Pre-processing
  - Spelling correction?
  - Word type or lemma?

    Lemmatization *write*, *wrote*, *written* → *write*

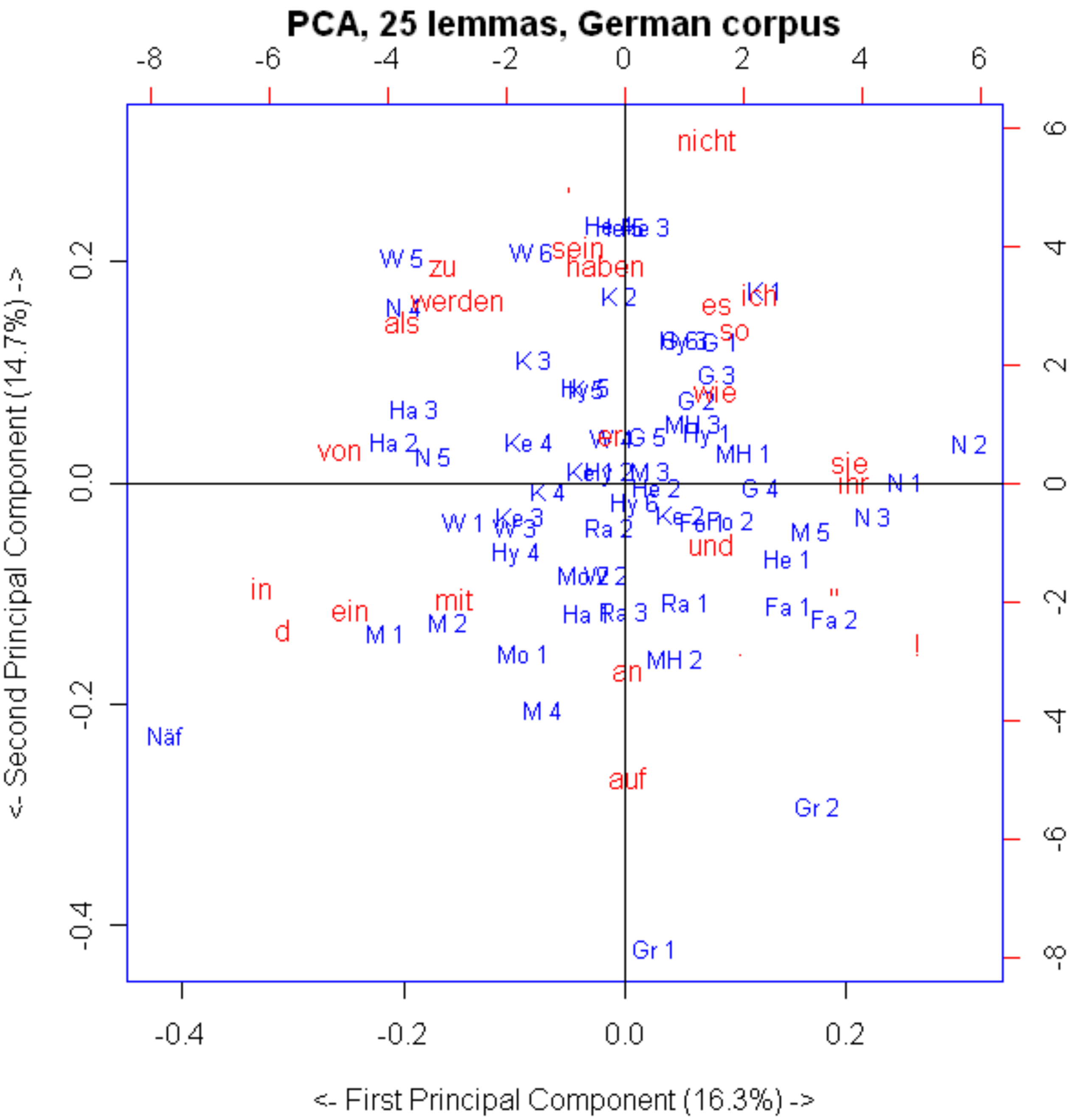
    der, *das*, *die* → *d*

    aimes, *aimons* → *aimer*

German Corpus

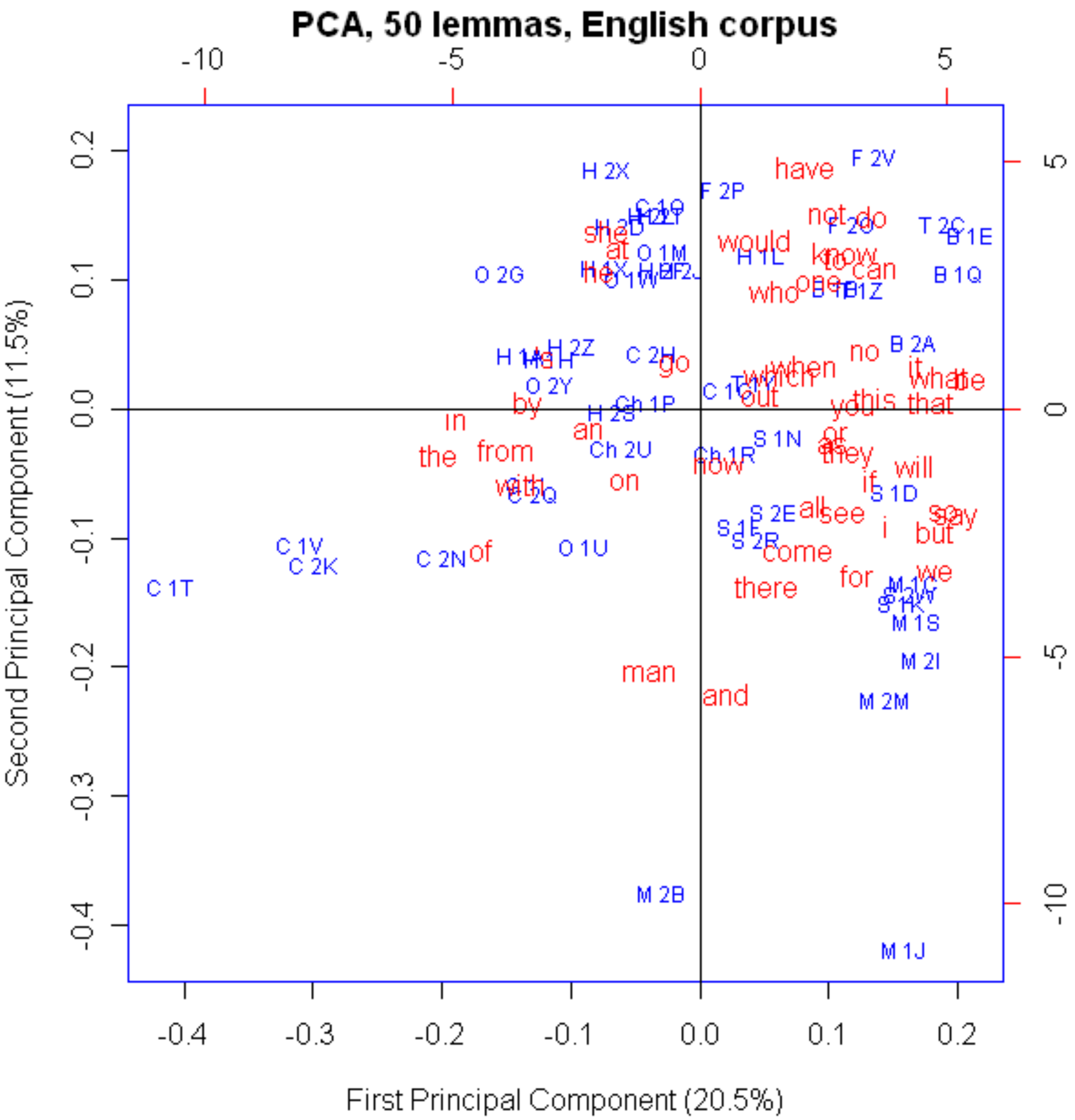
Author	Title 1	Title 2	Title 3
Goethe	*Die Wahlverwandschaften*	*Die Leiden des jungen Werther*	*Wilhelm Meisters Wanderjahre*
Heyse	*L'Arrabbiata*	*Beatrice*	*Der Weinhüter von Meran*
Fontane	*Unterm Birnbaum*		
Nietzsche	*Also Sprach Zarathustra*	*Ecce Homo*	
Hauptmann	*Bahnwärter Thiel*	*Bahnwärter Thiel*	
Falke	*Der Mann im Nebel*		
H. Mann	*Flöten und Dolche*	*Der Vater*	
T. Mann	*Der Tod in Venedig*	*Tonio Kroeger*	*Tristan*
Kafka	*Die Verwandlung*	*In der Strafkolonie*	
Wassermann	*Caspar Hauser*	*Der Mann von vierzig Jahren*	*Mein Weg als Deutsche und Jude*
Hesse	*Knulp*	*Siddhartha*	
Graf	*Zur Freundlichen Erinnerung*		

PCA



German
25 lemmas
60 text
excerpts

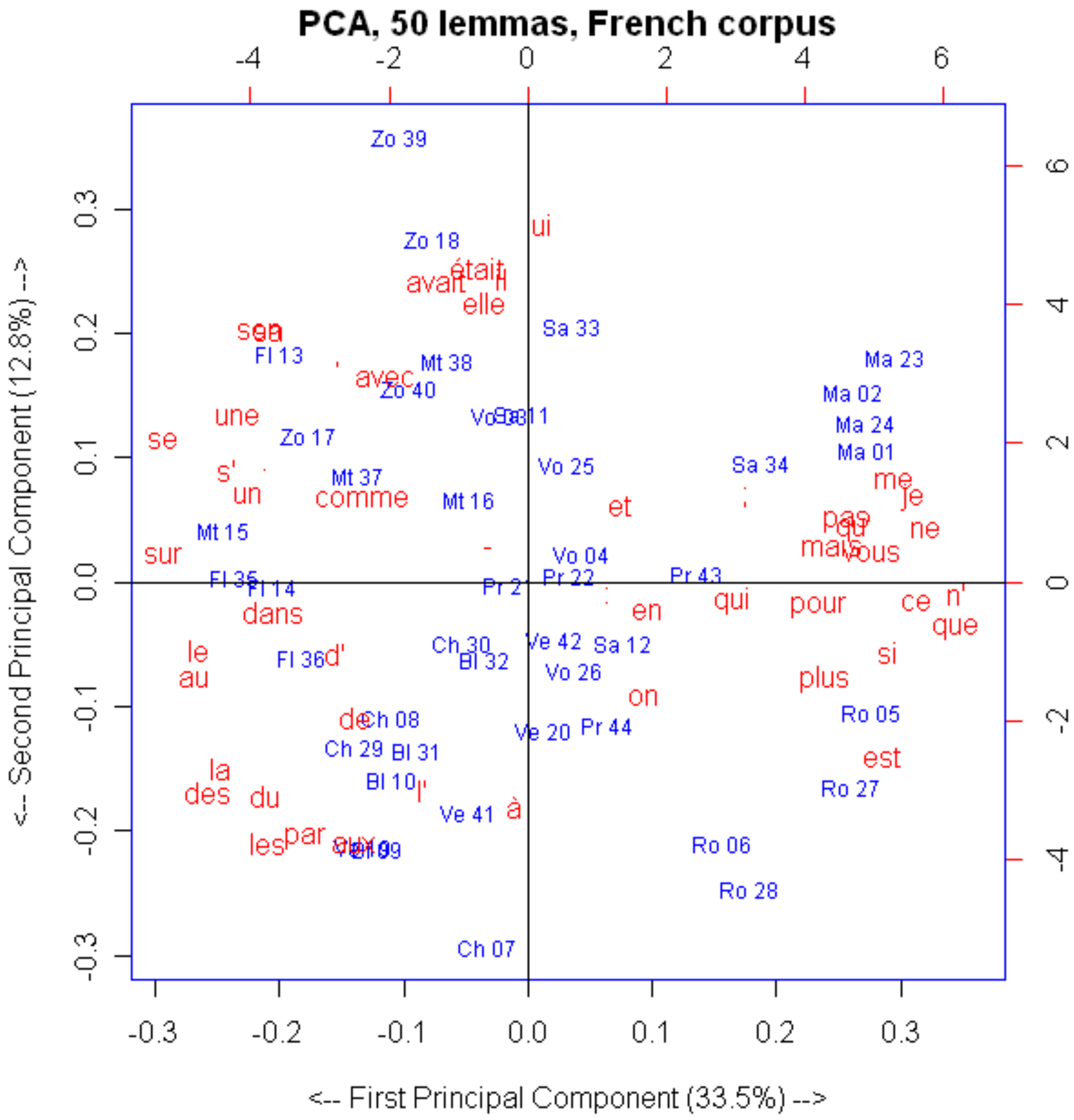
PCA



English
50 lemmas
52 excerpts

PCA

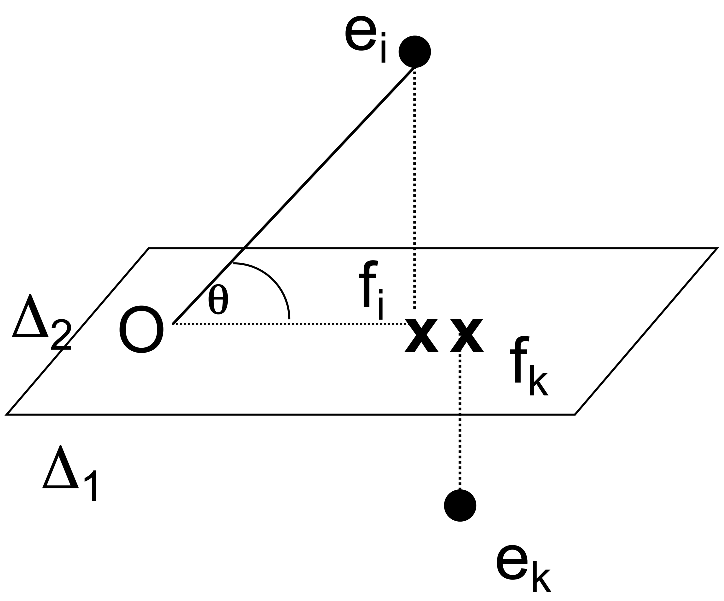
French
50 lemmas
44 text excer



Principal Component Analysis

Visual and real distance.

Having two points f_i and f_k close together in the PC1 and PC2 plan does not mean that the corresponding e_i and e_k points are also close together.



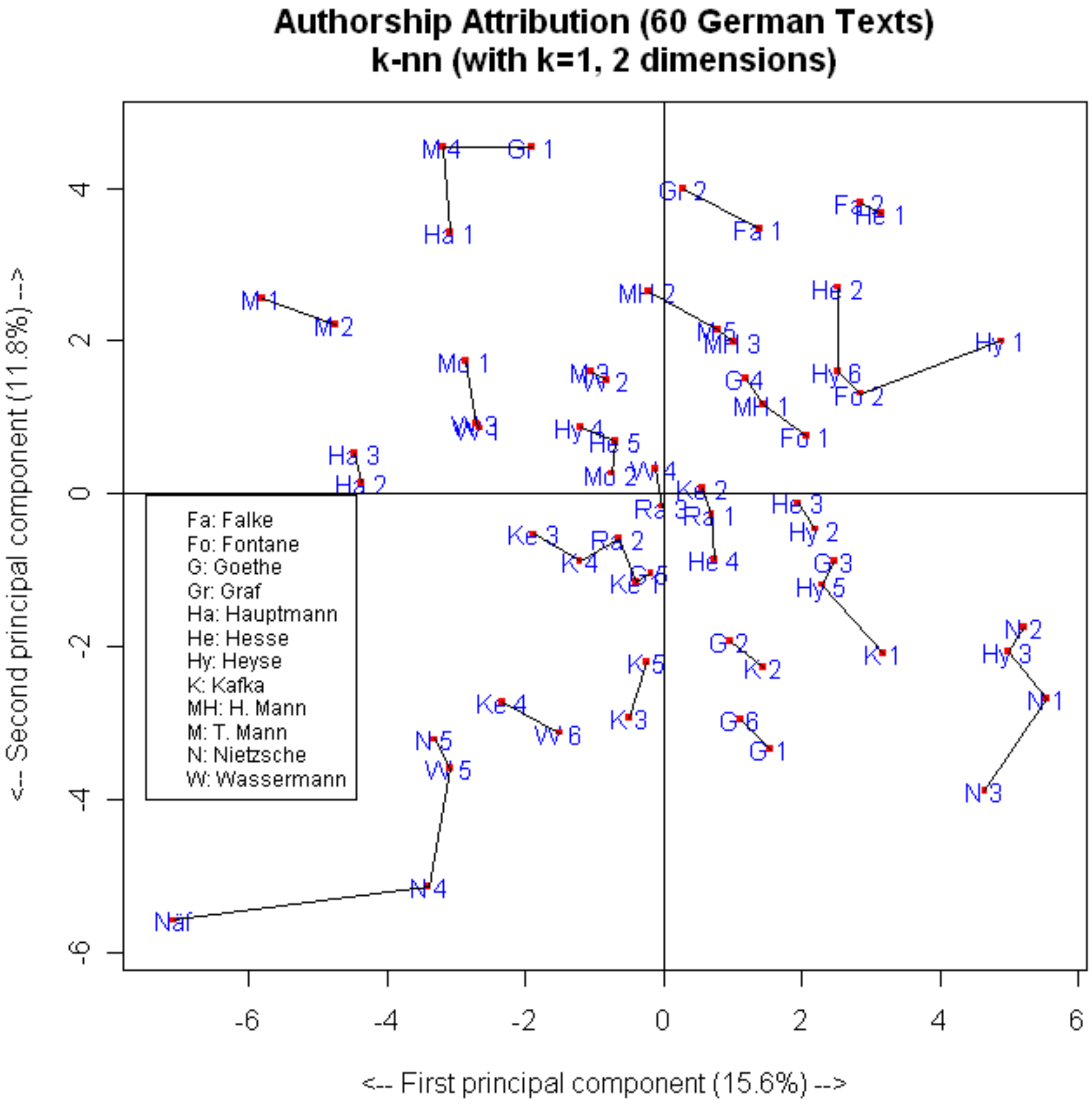
PCA could be useful in your context,

- to visualize
- to synthesis your data!
- some hints about the style

Nearest Neighbour

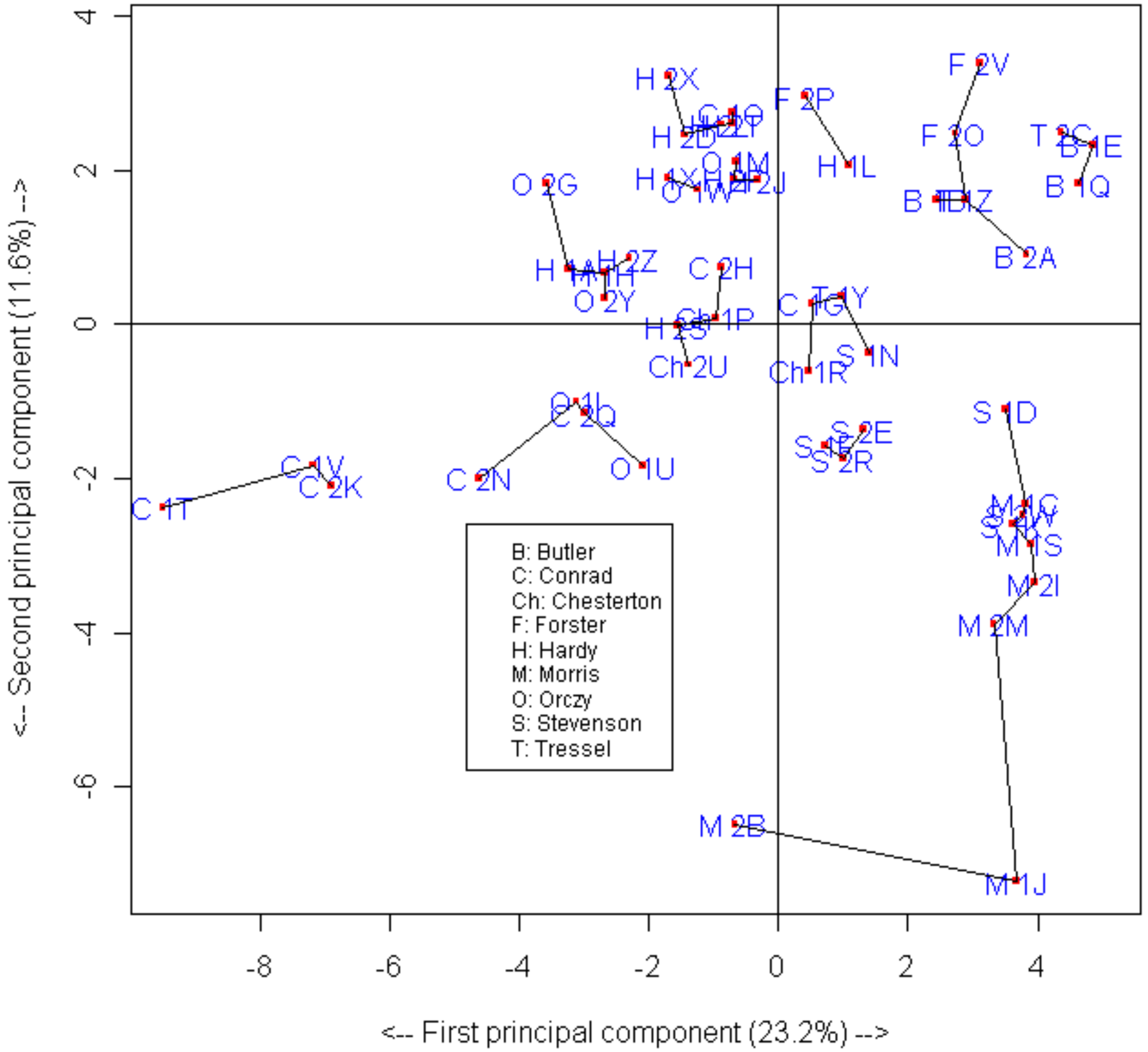
- Learning is just storing the representations of the training examples (all but not D_x)
- Testing instance D_x:
  - Compute similarity between D_x and all other examples
  - Assign D_x the category of the most similar example (1-NN)
- Does not explicitly compute a generalization or category prototypes
- Nearest neighbor method depends on a similarity (or distance) metric

PCA & NN



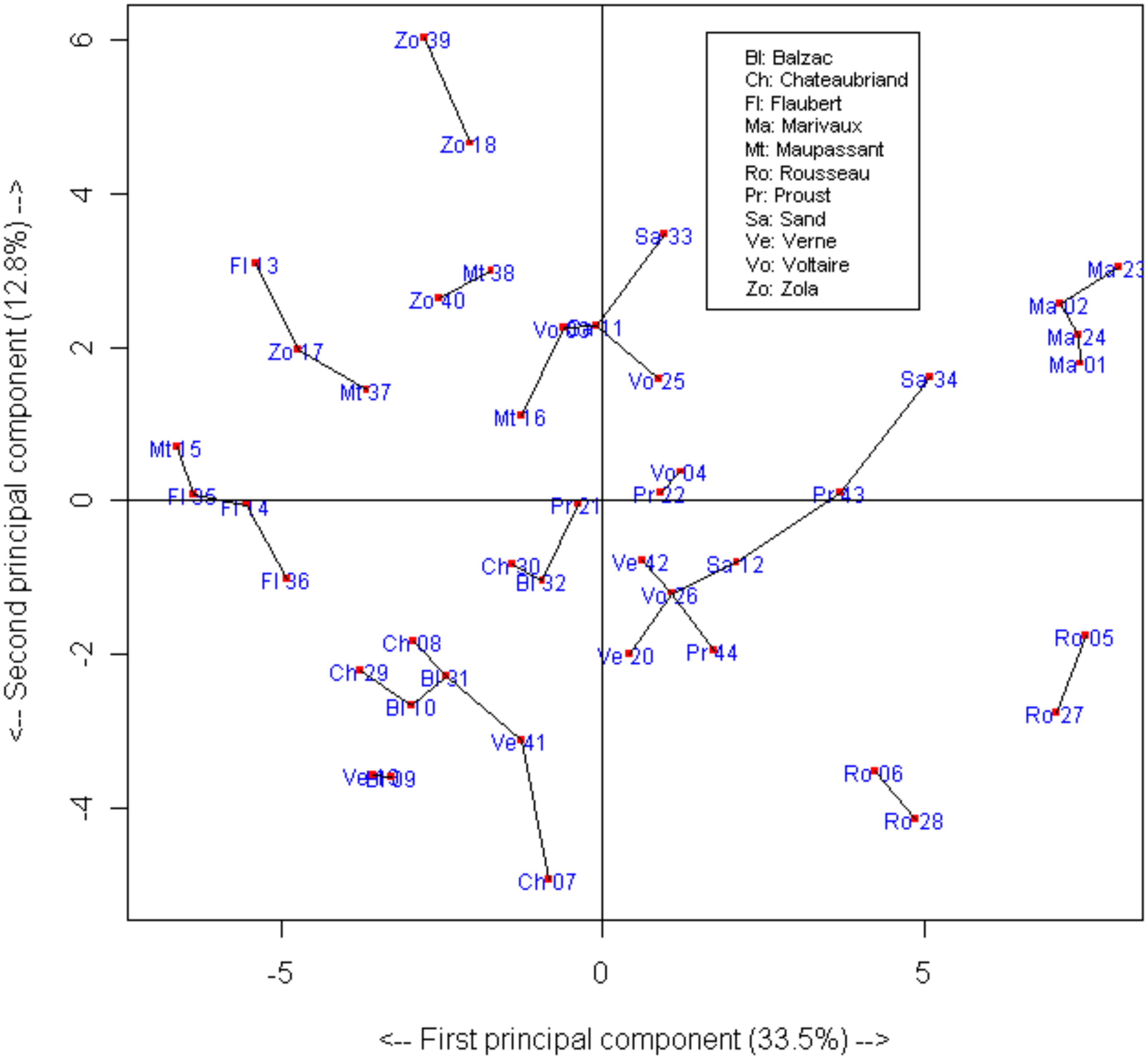
German
50 lemmas
60 excerpts

PCA & NN



English
50 lemmas
52 excerpts

PCA & NN



French
50 lemmas
44 text
excerpts

Evaluation

English Corpus, 52 text excerpts (~10 000 tokens), 9 authors

French Corpus, 44 texts excepts (~10 000 tokens), 11 authors

German Corpus, 59 texts excepts (~10 000 tokens), 15 authors

	English	**French**	**German**
PCA, 2 axes, 50 lemmas	36.5%	31.8%	30.5%
PCA, 5 axes, 50 lemmas	86.5%	68.2%	63.7%
PCA, 2 axes, 100 lemmas	57.7%	54.6%	39.0%
PCA, 5 axes, 100 lemmas	**92.3%**	**70.4%**	**66.1%**

Burrows' Delta

- Based on on the *n* most (*n* = 150) frequent words (+ POS for some types such as *to*, *in,* and expand others) "frequency-hierarchy for the most common words in a large group of suitable texts" (p. 269)
- Compute a Z-score value for each word
  - for each word type w_i , i = 1, …, *n* in a sub-corpus *D,* compute the relative frequency rf_{Di} (in ‰)
  - μ_i mean in the reference corpus
  - σ_i standard deviation

$$Z(w_{Di}) = \frac{rf_{Di} - \mu_i}{\sigma_i}$$

Burrows, J. F. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3), 267-287.

Burrows' Delta

First compute the author profile: sum the frequencies

	G1	G3	N25	N27	M39	M40	K42	K43	Näf
d	665	775	573	894	681	836	758	775	1162
.	345	254	267	318	348	398	351	363	362
und	258	307	323	148	443	473	197	201	183
sein	219	276	258	262	327	262	270	288	178
ich	172	426	203	309	98	48	220	151	1
in	122	133	63	182	177	183	95	124	296
nicht	105	97	128	107	81	52	152	130	66
werden	74	54	35	81	39	44	85	66	85

Burrows' Delta

	G	N	M	K	Näf
d	1440	1467	1517	1533	1162
.	599	585	746	714	362
und	565	471	916	398	183
sein	495	520	589	371	178
ich	598	512	146	371	1
in	255	245	360	219	296
nicht	202	235	133	282	66
werden	128	116	83	151	85

Relative frequencies: divide by the sum (indep. size)

Burrows' Delta

Compute the mean (μ_i), standard deviation (σ_i), then the Z score

	G	N	M	K	Näf	μ	σ
d	0.336	0.353	0.338	0.363	0.498	0.378	0.068
.	0.140	0.141	0.166	0.169	0.155	0.154	0.014
und	0.132	0.113	0.204	0.094	0.078	0.124	0.049
sein	0.116	0.125	0.131	0.132	0.076	0.116	0.023
ich	0.140	0.123	0.033	0.088	0.000	0.077	0.59
in	0.060	0.059	0.080	0.052	0.127	0.075	0.031
nicht	0.047	0.057	0.030	0.067	0.028	0.046	0.017
werden	0.030	0.028	0.018	0.036	0.036	0.030	0.07

Burrows' Delta

- Distance between two sub-corpora D (doubtful) and D' (known)

  If Δ is small, D and D' are written by the same author.

$$\Delta(D, D') = \frac{1}{n} \sum_i^n |Z(w_{Di}) - Z(w_{D'i})|$$

- Modification suggested (Hoover, 2004)
  - *n* must be greater than 150 (e.g., 800)
  - ignoring personal pronouns
  - culling at 70% (words for which a single text supplies more than 70% of the occurrences)

Hoover, J. F. (2004). Delta Prime? *Literary and Linguistic Computing*, 19(4), 477-495.

Burrows' Delta

Compute the distance with an unknown text

	G	N	M	K	Näf	test
Δ dist.	**6.25**	**6.22**	**9.37**	**5.06**	**7.67**	
d	-0.607	-0.356	-0.584	-0.219	1.765	0.879
.	-1.052	-0.975	0.876	1.082	0.070	0.330
und	0.154	-0.224	1.630	-0.619	-0.941	-0.821
sein	-0.021	0.397	0.651	0.688	-1.716	0.129
ich	1.062	0.787	-0.747	0.186	-1.289	-0.393
in	-0.521	-0.538	0.153	-0.773	1.679	-0.639
nicht	0.089	0.652	-0.958	1.255	-1.037	0.046
werden	0.027	-0.242	-1.545	0.832	0.928	0.497

Evaluation

English Corpus, 52 text excerpts (~10 000 tokens), 9 authors

French Corpus, 44 texts excepts (~10 000 tokens), 11 authors

German Corpus, 59 texts excepts (~10 000 tokens), 15 authors

	English	French	German
Delta, 50 word types	96.4%	86.4%	79.7%
Delta, 100 word types	**98.1%**	81.8%	**84.7%**
Delta, 150 word types	96.2%	**90.9%**	**84.7%**
PCA, 5 axes, 100 lemmas	92.3%	70.4%	66.1%

Z Score

The absolute frequency is ignored in Burrows' Delta rule.

	McCain'08	rest	C
"Bush"	26	398	424
not "Bush"	154,339	474,331	628,670
	154,365	474,729	629,094

- Prob["Bush" in C] = 424/629,094 = 0.000674.
- *n'* = 154,365
- We expect in McCain'08 $n' \cdot \mathrm{Prob}[\omega] = 104.04$
- Z score ("Bush" in McCain'08) = -7.65

Z Score

The Z score values for some very frequent German lemmas

between -2 and 2, normal usage

negative value → under-used, positive value → over-used

Lemma	Goethe	Kafka	Nietsche	Hesse	T. Mann
d	**-3.66**	**3.39**	-0.75	**-5.80**	**3.31**
.	**-4.20**	**-2.76**	**-4.66**	0.54	-0.44
und	**-2.79**	**-5.51**	0.57	**2.42**	**4.91**
sein	-1.13	-0.01	0.72	**4.14**	1.58
ich	**4.76**	**-4.66**	**7.51**	1.55	**-8.07**
nicht	0.67	**3.60**	0.40	1.23	**-2.60**

Z Score: A. Näf vs. Others

The over-used terms are *Schüler*, *insgesamt*, *Ergebnis*, *Klasse*, *Resultat*, *Schuljahr*, *Schülerin*, …

Lemma	Goethe	Kafka	Nietsche	Hesse	T. Mann	A. Näf
d	**-3.66**	**3.39**	-0.75	**-5.80**	**3.31**	**13.83**
.	**-4.20**	**-2.76**	**-4.66**	0.54	-0.44	-1.00
und	**-2.79**	**-5.51**	0.57	**2.42**	**4.91**	**-8.10**
sein	-1.13	-0.01	0.72	**4.14**	1.58	**-5.70**
ich	**4.76**	**-4.66**	**7.51**	1.55	**-8.07**	**-13.34**
nicht	0.67	**3.60**	0.40	1.23	**-2.60**	**-2.53**

Z Score

- We have a Z score for each term t_i in a document D_j

$$Zscore(t_{ij}) = \frac{a - (n' \cdot Prob[t_{ij}]}{\sqrt{n' \cdot Prob[t_{ij}] \cdot (1 - Prob[t_{ij}])}}$$

- When comparing two texts, considering all Z scores

$$Dist(D_j, D_k) = \frac{1}{m} \sum_i^m \left(Zscore(t_{ij}) - Zscore(t_{ik})\right)^2$$

Evaluation

English Corpus, 52 text excerpts (~10 000 tokens), 9 authors

French Corpus, 44 texts excepts (~10 000 tokens), 11 authors

German Corpus, 59 texts excepts (~10 000 tokens), 15 authors

	English	**French**	**German**
Z score	**100%**	**100%**	**84.7%**
Delta, 150 word types	96.2%	90.9%	**84.7%**
PCA, 5 axes, 100 lemmas	92.3%	70.4%	66.1%

Conclusion

- Authorship attribution
  - More than only literature novels / historical documents
  - Mainly based on the vocabulary (and the occurrence frequencies)
- Various approaches
  - Single measure
  - Multivariate analysis (PCA)
  - Text categorization approach (machine learning)
- Next step
  - Shorter text excerpts, larger number of text excerpts and authors
  - Uncertainty
  - “Le style c’est l’homme”, Comte de Buffon
  - Selection and weighting of the features
  - Better classifier
  - Other medium

English Corpus

Nb	Author	Short Title	Title
4	**Butler**	*Erewhon*	*Erewhon revisited*
3	**Chesterton**	*Man who was*	*Man who was Thursday*
4	**Conrad**	*Almayer*	*Almayer's Folly*
4	**Conrad**	*Lord Jim*	*Lord Jim*
3	**Forster**	*Room with view*	*A Room with a View*
3	**Hardy**	*Jude*	*Jude the Obscure*
3	**Hardy**	*Madding*	*Far from the Madding Crowd*
4	**Hardy**	*Well beloved*	*The Well-Beloved*
2	**Hardy**	*Wessex Tales*	*Wessex Tales*
2	**Morris**	*Dream of JB*	*A Dream of John Ball*
4	**Morris**	*News*	*News from Nowhere*
3	**Orczy**	*Elusive P*	*The Elusive Pimpernel*
3	**Orczy**	*Scarlet P*	*The Scarlet Pimpernel*
3	**Stevenson**	*Ballantrae*	*The Master of Ballantrae*
4	**Stevenson**	*Catriona*	*Catriona*
3	**Tressel**	*Ragged TP*	*The Ragged Trousered Philanthropists*

French Corpus

Author	**Title 1**	**Title 2**
Marivaux	*La Vie de Marianne*	*Le Paysan parvenu*
Voltaire	*Zadig*	*Candide*
Rousseau	*La nouvelle Héloïse*	*Emile*
Chateaubriand	*Atala*	*Vie de Rancé*
Balzac	*Les Chouans*	*Le cousin Pons*
Sand	*Indiana*	*La Mare au Diable*
Flaubert	*Madame Bovary*	*Bouvard et Pécuchet*
Maupassant	*Une Vie*	*Pierre et Jean*
Zola	*Thérèse Raquin*	*La Bête humaine*
Verne	*De la Terre à la Lune*	*Le Secret de Wilhelm Storitz*
Proust	*Du côté de chez Swann*	*Le Temps retrouvé*