

Correlaciones inter-lingüísticas Διαγλωσσική

απόδοση συγγραφέων

والارتباطات اللغوية Cross-

linguistic correlations *Correlações inter-linguísticas*

ຄວາມສໍາພັນ ລະຫວ່າງພາສາ in lexical complexity:

An approach to cross-linguistic authorship attribution

Patrick Juola

Duquesne University, USA

George Mikros

National and Kapodistrian

University of Athens, Greece



# Authorship attribution

- Classic problem in scholarship, including literature, forensic/legal, and historical scholarship
- Given a document, who wrote it?
  - If you can't tell me that, can you tell me something about who wrote it?
  - Increasing research area with substantial body of work



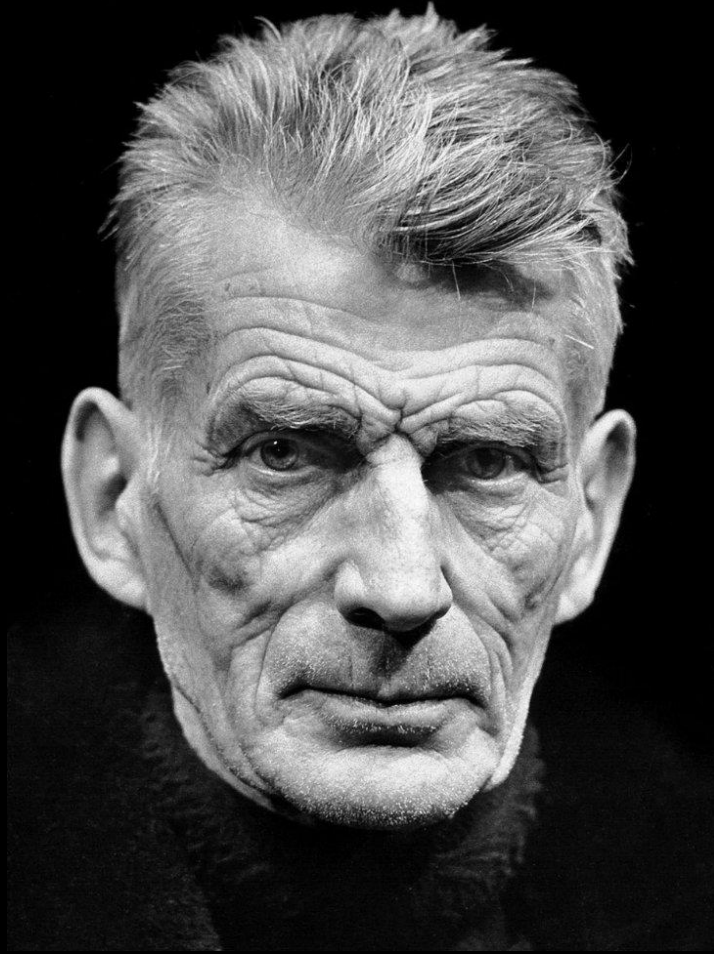
# Matching KD to QD

- I want the KD to match the QD as closely as possible
  - Genre, tone, date, subject,
- I also want
  - ... a unicorn
  - ... that flies and grants wishes
- How do I “match” a suicide note?
  - ... or a document across languages?



# Documents across languages

- J.K. Rowling was easy. Everything she wrote was in English.
- How about Samuel Beckett?
- Published in French and English, also fluent in Italian
  - See also Nabokov, *inter al.*



# Documents across languages

- Joseph Conrad spoke
  - ... Polish natively, but also
    - Latin
    - German
    - Greek
    - French
  - ... but preferred to write in English



# Application: R.A.M.P.

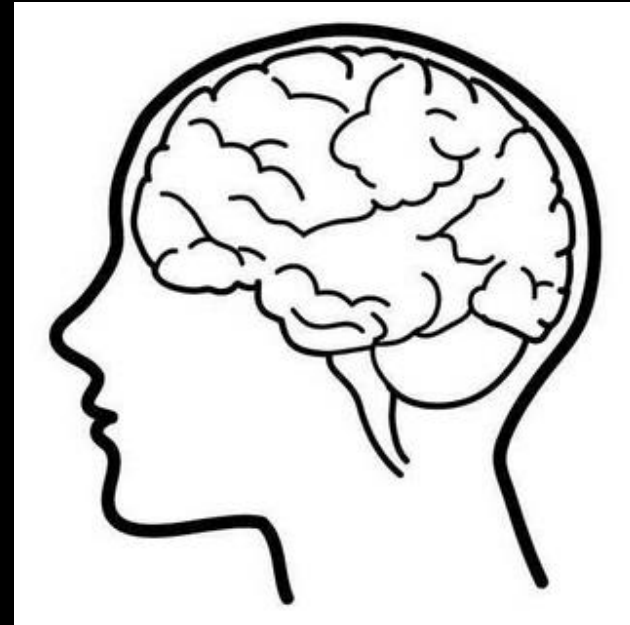
- Russian Anonymous MarketPlace
- “Deep Web” drug (&c.) e-commerce
- Transactions are a) in Russian, b) anonymous, c) often highly illegal
- Law enforcement really wants to trace online identities to meatspace, but most of the surface Web is in English.
  - Need to map Russian posts to English





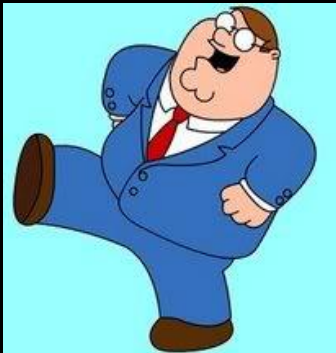
# What is the same?

- Need to identify features common across genres (and languages)
- “Words” and “characters” won’t work.
- But the authorial “mind” is the same!



# Can we identify “minds”?

■ The paradigmatic and systematic utilization of sesquipedalian lexical items can be an informative element of individual and idiosyncratic patterns of linguistic variation



Or, some people use big words





# Cross-linguistic Feature Set

- Need to identify features that are not tied to language
- Some candidates include
  - Word lengths
  - Utterance lengths
  - Type-Token Ratio
  - *hapax legomena*
  - Use of names
  - Use of references
  - Social conventions (e.g #hashtags on Twitter)



# Prior Work

- Hand-identified bilinguals (Sp/En) on Twitter
- Lg. of each tweet checked by machine
- 20-664 tweets/language/user
- Sample size controls

User	English (reliable)	Spanish (reliable)
S01	51	263
S02	49	61
S03	313	25
S04	116	218
S05	18	38
S06	280	146
S07	140	94
S08	167	654
S09	62	60
S10	47	103
S11	468	664
S12	157	127
S13	35	161
S14	20	38



# Our attempt at features

- Various measures of “linguistic complexity”
  - Word length, type-token ratio, % hapax legomena, Entropy, Yule’s K, Popescu’s *R1 inter al.*
  - Mostly taken from QUITA package ([www.quitaonline.com](http://www.quitaonline.com))



# Does <mind> correlate?

Yes.

For gearheads:

Index	<i>r (untrunc)</i>	<i>p-val (untrunc)</i>	<i>r (first 200)</i>	<i>p-val (first 200)</i>
<b>TTR</b>	0.2690	p>0.1(n.s.)	0.7431489	0.002321(**)
<b>Hapax %</b>	0.2750575	p>0.1(n.s.)	0.7142336	0.002054 (**)
<b>Entropy</b>	0.3307389	p>0.1(n.s.)	0.7059129	0.002392 (**)
<b>Lambda</b>	0.3925311	0.08253 (n.s.)	0.6937589	0.002961(**)
<b>Redundancy</b>	0.3013017	p>0.1 (n.s.)	0.6924124	0.00303 (**)
<b>Popescu's R1</b>	0.7174944	0.001933 (**)	0.3504229	p>0.1 (n.s.)
<b>Yule's K</b>	0.4209762	0.06694 (n.s.)	0.5906128	0.01308 (*)
<b>RR</b>	-0.02707195	p>0.1 (n.s.)	0.5633507	0.01796 (*)
<b>RRmc</b>	0.6671255	0.004576 (**)	0.2678785	p>0.1 (n.s.)
<b>L</b>	0.6394239	0.006903 (**)	0.6584157	0.00523 (**)
<b>Adj. Mod.</b>	0.5986507	0.01185 (*)	0.485988	0.03904 (*)
<b>G</b>	0.5021349	0.03365 (*)	0.6544689	0.005549 (**)
<b>Curve Length R</b>	0.5021349	0.05499 (n.s.)	0.4983632	0.03486 (*)

# Does this generalize?

- Gosh, Patrick, can't you get *anything* right?
  - Twitter is not really representative of ordinary writing
  - Spanish/English are typologically close
  - The sample size is too small
  - Where are the topic controls?

» Ooops?!







# George to the rescue!

- English/Modern Greek corpus
  - 100 subjects, 2 topics, 2 languages
    - Topic 1 : Personal (“Happy memories”)
    - Topic 2 : Argumentative (“Immigration”)
  - All writers native Greek, prof. Eng.
  - Roughly 1000-3000 words (English)
  - All documents in “standard” written language.
- What happens when we look at lg. complexity now?



# Our measures

- Again, taken from QUITA
  - Greek alphabet/language is not a problem!
- 37 individual measures, including type/token ratio, average word length, % hapax legomena, &c.
- Measured for each essay, then correlated across 100 Ss.



# (Partial) Correlation results

## Strong correlations on same topic

<i>Index</i>	<i>r</i> (Topic 1)	<i>r</i> (Topic 2)	<i>Index</i>	<i>r</i> (topic 1)	<i>r</i> (topic 2)
TTR	0.818616007	0.784786047	G	0.751217331	0.777963375
h	0.4778307	0.483425942	Yule's K	0.375775184	0.283666006
Entropy	0.806372567	0.717849433	Hapax %	0.807895819	0.752252585
AWL	0.832311407	0.856971495	L	0.865193214	0.854336884
R1	0.604693819	0.419052444	Writers' View	0.210550065	0.276745255
RR	0.363035899	0.288639524	CL_Rindex	0.220754857	0.18364659
RRmc	0.296769035	0.288831596	DL	0.231716829	0.272513487
Lambda	0.749918083	0.753646307	D_H	0.501308357	0.499486779
AdjMod	0.662598756	0.605278122	LD	0.536039704	0.530153323


# ... even on different topics

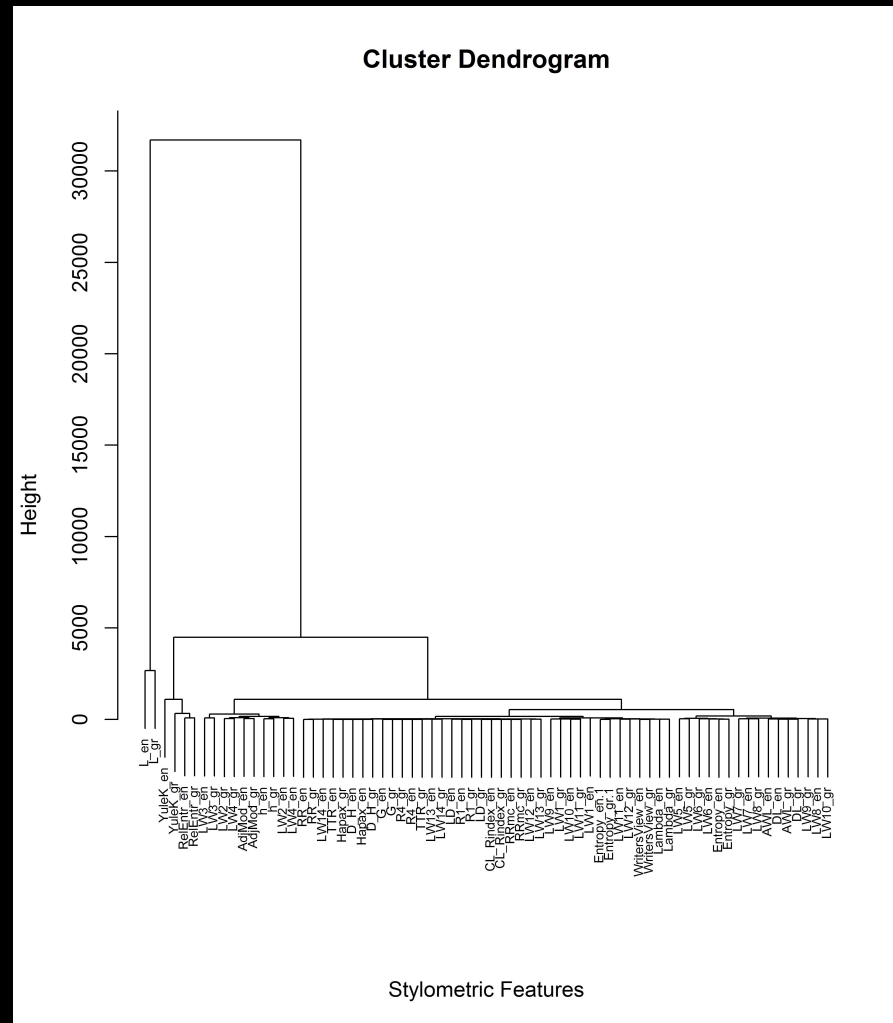
## Reasonable correlations on different topics

<i>Index</i>	<i>r</i> (Topic 1->2)	<i>r</i> (Topic 2->1)	<i>Index</i>	<i>r</i> (topic 1->2)	<i>r</i> (topic 2->1)
TTR	0.530131717	0.495351349	G	0.545595344	0.504398647
h	0.376987718	0.35512703	Yule's K	0.215560384	0.060845028
Entropy	0.374647315	0.409135171	Hapax %	0.504153445	0.446729209
AWL	0.05536855	0.106453929	L	0.561028393	0.601818658
R1	0.322406798	0.130281062	Writers' View	0.187018449	0.241829074
RR	0.198938717	0.060061702	CL Rindex	0.16414749	0.172832719
RRmc	0.254576958	0.050050319	DL	0.141512942	-0.033222034
Lambda	0.446972084	0.423402726	D H	0.317073352	0.180292177
AdjMod	0.281611794	0.28180575	LD	0.30719122	0.086930288



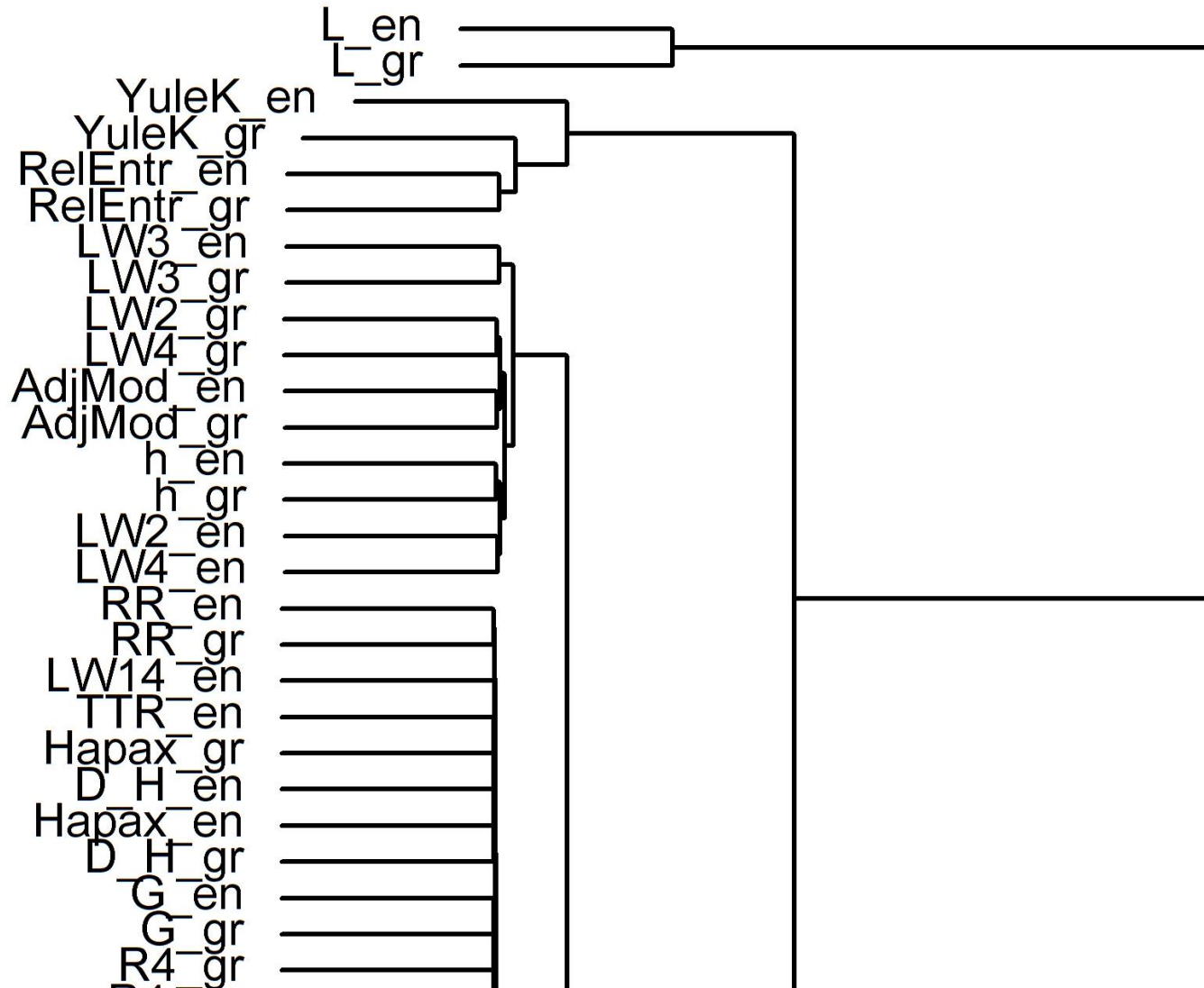
# Cluster analysis

 Some degree of independence among features





0 5000 10000



St

# Analysis of first 500 words

■ Still generally good correlations

<i>Index</i>	<i>r</i> <i>(Both topics)</i>	<i>Index</i>	<i>r</i> <i>(Both topics)</i>
TTR	0.733810585	G	0.67143317
<i>h</i>	0.079145105	Yule's K	0.67143317
Entropy	0.581922258	Hapax %	0.697098723
AWL	0.914778143	L	0.703526446
R1	0.37815215	Writers' View	-0.156974712
RR	0.20575165	CL_Rindex	-0.062065421
RRmc	0.182504193	DL	0.25158039
Lambda	0.700898362	D_H	0.293979624
AdjMod	0.453160609	LD	0.516198927



# Discussion of findings

- Most correlations positive, meaning
  - The more complex your writing in Greek, the more complex in English!
- Most correlations are significant, some extremely so.
  - Confirmed: Spanish not a fluke
- Topic dissimilarity reduces but does not hide correlation



# Next: Can we use this?

- In other words, is this finding useful?
- Implicit 18-dimensional vector space
- Use this space to classify writings
  - Previous work on Spanish suggests possible but not accurate enough to be useful, but see experimental issues
  - Work in progress



# Discussion

- What linguistic “features” are we looking at here?
- What other types of features might be both transferrable and easy to extract?
- Can anyone else contribute new language pairs? (We’re always looking for new heroes!)





Nevertheless....



- Dank u vel
- Ευχαριστώ πολύ
- Děkuju rěkně
- Dziękuję bardzo
- Muito obrigado
- Thank you very much
- Спасибо
- Asanteni
- Çok teşekkür ederim



# A final advertisement



SIX SEPTEMBERS

## Mathematics for the Humanist

Patrick Juola and Stephen Ramsay

- New book
- High-level math for smart humanists
- <http://digitalcommons.unl.edu/zeabook/55/>
- Free for e-book, nominal cost for soft-cover

