TABULA RASA: Trusted Biometrics under Spoofing Attacks
CSSC, Rome, 10th May 2012

Linguistic profiling as an aspect of Active Authentication

Some challenges

Claire Hardaker
Forensic/corpus linguist, UCLan/Lancaster University

Support and challenge

- I **support** R's view that linguistic profiling
  - is a viable method of individuating users
  - must be used alongside other modalities
- However, I **challenge** with regards to
  - issues faced at various linguistic levels
  - how 'optimal' chosen levels are or can be
- I finish by tentatively proposing 'perfect' profilers, data, and features

Scope of the challenge

- My challenge looks at **computer-mediated communication** (interaction *via* device)
- I marginally consider **human-computer interaction** (interaction *with* device)
- (I defer the supporting/challenging of non-linguistic modalities to those with more knowledge/expertise)

Issue #1:
Features and focus

Consistency and distinctiveness

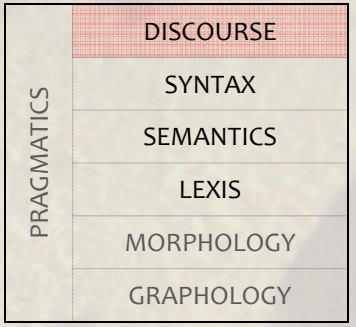
- **(Forensic/computational/stylometric) linguistic analysis alone cannot say *who* wrote a text**
- It works best as a support/challenge to other evidence, but to do so…
  - A needs to **consistently** use certain features
  - A's features need to be **distinctive** from B/C/D's
  - we must have a **robust comparison corpus**
- n.b. most (all?) current linguistic analysis is on a 'final product', not real-time

Qualitative versus quantitative

QUALITATIVE ***what kind of person wrote this?*** *e.g. sociolinguistics, discourse analysis, forensic linguistics*	**QUANTITATIVE** ***did person A write this?*** *e.g. corpus linguistics, computational linguistics, stylometrics*
Rich in user information	Lean in user information
(Typically) suited to small datasets	(Typically) needs large datasets
Non-generalisable	Generalisable
Inter-rater reliability issues	No inter-rater reliability issues
Features (typically) chosen afterwards—subjective	Features (typically) chosen beforehand—principled

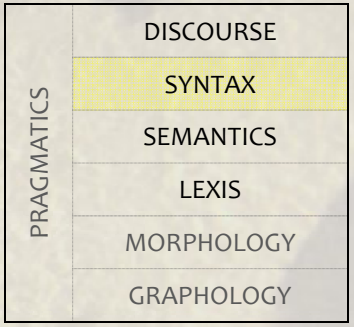
Issue #2:
Linguistic levels

Context (discourse)



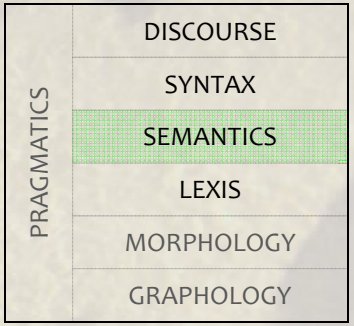
- Linguistic choice depends on many factors:
  - **Domain** SMS, email, article, etc.
  - **Formality** courtroom hearing, casual chat, etc.
  - **Topic** disciplinary, gossip, etc.
  - **Purpose** persuade, inform, threaten, etc.
  - **Author** executive, model, moderator, declared
  - **Audience** hierarchy, relationship, later reuse
  - **Mode** dictated, written, typed
  - **Device** length, (lack of) editing, lexis/errors

Grammar (syntax)



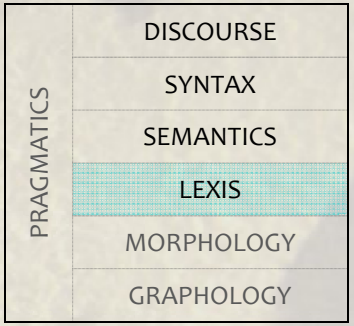
- Sentence/word/punctuation metrics can modestly individuate users, **but** stylometry software[1] works best with closed sets
- Syntactic habits/choices can modestly individuate users, **but** taggers[2] rely on standard syntax and spelling
- (Software[4] can detect/fix spelling variants, but I'm not aware of a syntax 'fixer')

Meaning (semantics)



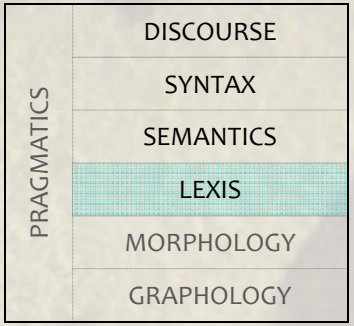
- Semantic taggers[2,3] can identify semantic fields, which can *help* with categorising texts based on their domain, topic, register, etc.
- This in turn helps to create more robust comparison (sub-)corpora
- **Problem**: semantics is subjective and less amenable to quantitative comparison

Vocabulary (lexis)



- The English open class lexicon (vb, nn, av, aj) is huge (~600k-1m) and changes *very* quickly
- Individual lexicons are a result of our lives and vary in richness/scope (~30k-75k)
- **Problem:** OCL choice is *heavily* influenced by discourse-level factors
  - Quantitative OCL comparisons *must* be cautious
  - But qualitative lexicon analysis can be insightful!

Vocabulary (lexis)



- The English closed class lexicon (pn, dt, pp, cj, ax, md) is tiny (~450), changes v. slowly
- Our CCL use is not usually influenced by discourse-level factors, so we can compare results across text-types
- **Problem**: informality/brevity especially in CMC typically affects CCL, e.g.:

Just sending email. Meet you there?

In summary...

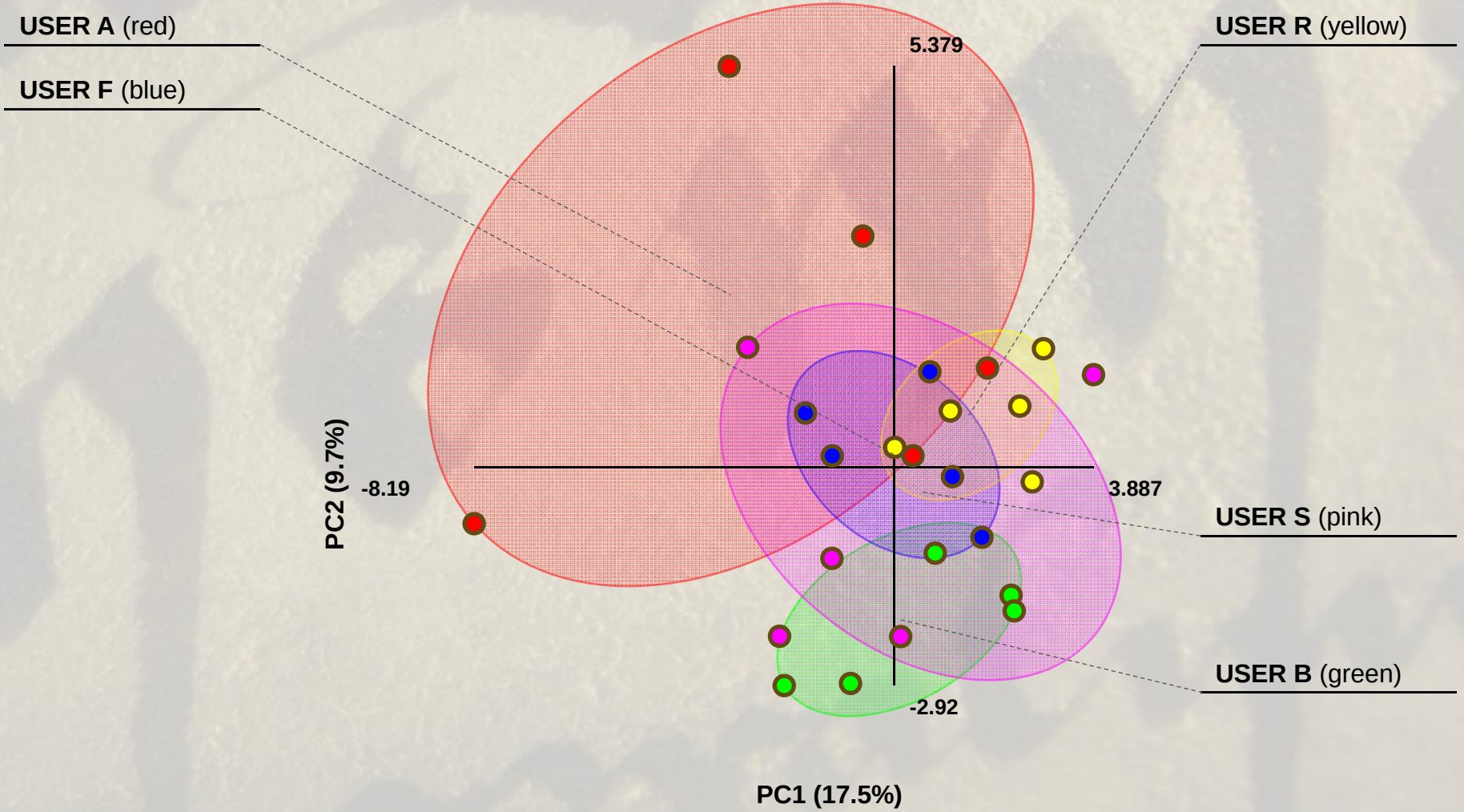
A ‘perfect profiler’…

- A ‘perfect’ linguistic profiler would…
  - derive metadata (e.g. from To: CC: BC: fields)
  - parse/tag/count (i.e. syntax, semantics, lexicon)
  - create meaningful, robust, ‘clean’ sub-corpora
  - quantitatively analyse these corpora
  - allow further qualitative analysis if necessary
  - and produce results that integrate well (as a support or challenge) with other biometrics

‘Perfect data’…

- *Not* perfect, but… emails as training data?
  - contains meta-data (i.e. sender/receiver, etc.)
  - typically one typist (n.b. *can* be dictated)
  - analysis can be triggered by clicking ‘send’
  - wealth of data already stored on servers
    - post-process existing data
    - sub-corpora creation straightforward

‘Perfect features’…



Thank you! ☺

Software

1. *Signature*: http://www.philocomp.net/humanities/signature
2. *Wmatrix/USAS/CLAWS*: http://ucrel.lancs.ac.uk/wmatrix/
3. *WordNet*: http://wordnet.princeton.edu/
4. *VARD2*: http://ucrel.lancs.ac.uk/VariantSpelling/