CLiPS
Computational Linguistics & Psycholinguistics
University of Antwerp

Profiling in Practice

Walter Daelemans
CLiPS Computational Linguistics Group
walter.daelemans@uantwerpen.be

PAN @ CLEF

Toulouse, September 9, 2015

- Current AMiCA researchers @CLiPS:

- Other stylometry researchers currently @CLiPS:

- Current AMiCA researchers @CLiPS:



- Other stylometry researchers currently @CLiPS:



We have a postdoc vacancy for 4 years on a new BioNLP project! (Contact me if interested)

Plan of the talk

- Profiling: learning about the author through text
- Special challenges of social media text
- Application case study: AMiCA project
- New university spin-off company: Textgain

Richard Dawkins @RichardDawkins Mar 11

Islamists really going to town today. They'll gatecrash private dinners in revenge for our "forcing" Muslim women to sit next to us! WHAT?

Expand

Indignation
Derision
Anger

force

Who? we

What? sit

Who? muslim woman (plur)

Where? next to us

What? revenge

Why? gatecrash

Who? Islamist (plural) Religion

What? private dinner (plural) Parties

Richard Dawkins @RichardDawkins Mar 11

Islamists really going to town today. They'll gatecrash private dinners in revenge for our "forcing" Muslim women to sit next to us! WHAT?

Expand

On the basis of a set of tweets:

Profile

Man
Highly educated
60+
BrE native speaker
Introverted
…
= RD (authorship attribution)

Author identity detection is the limit case of profile detection

Computational Stylometry

- Writing style: A *combination* of *invariant* and *unconscious* decisions in language production at all linguistic levels (discourse, syntactic structures, lexical choice, …) *associated* with specific authors or author traits:
  - Age, gender, education level, native language, personality, emotional state, mental health, region, deception, (ideology, political conviction, religious beliefs,) …

https://watson-pi-demo.mybluemix.net/

Input Text

We need a minimum of 3500 words and ideally 6000 words or more to compute statistically significant results. See the science behind the service.

Ideally, the text should contain words we use in every day life relating to personal experiences, thoughts and responses. See usage guidence for details.

Choose Language: (•) English ( ) Spanish

Effective methods for evaluating the reliability of statements issued by witnesses and defendants in hearings would be extremely valuable to decision-making in Court and other legal settings. In recent years, methods relying on stylometric techniques have proven most successful for this task; but few such methods have been tested with language collected in real-life situations of high-stakes deception, and therefore their usefulness outside laboratory conditions still has to be properly assessed.
DeCour - DEception in COURt corpus - has been

449 words

Clear Analyze

Your Personality*

You are shrewd, skeptical and tranquil.

You are empathetic: you feel what others feel and are compassionate towards them. You are imaginative: you have a wild imagination. And you are philosophical: you are open to and intrigued by new ideas and love to explore them.

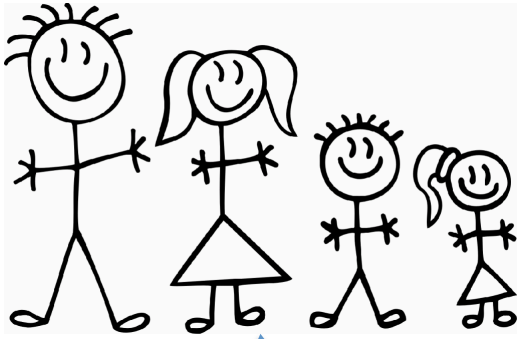
You are motivated to seek out experiences that provide a strong feeling of prestige.

You are relatively unconcerned with tradition: you care more about making your own path than following what others have done. You consider independence to guide a large part of what you do: you like to set your own goals to decide how to best achieve them.

**Compared to most people who participated in our surveys.*
***There were 451 words in the input. We need a minimum of 3,500, preferably 6,000 or more, to compute statistically significant estimates.*

Computational Stylometry as a *noisy channel* problem

Individual Variation → **Language Variation**



Language Psychology
Sociolinguistics

Spelling, punctuation, lexical choice, sentence structure, themes, tone, discourse structure, …

Profiling

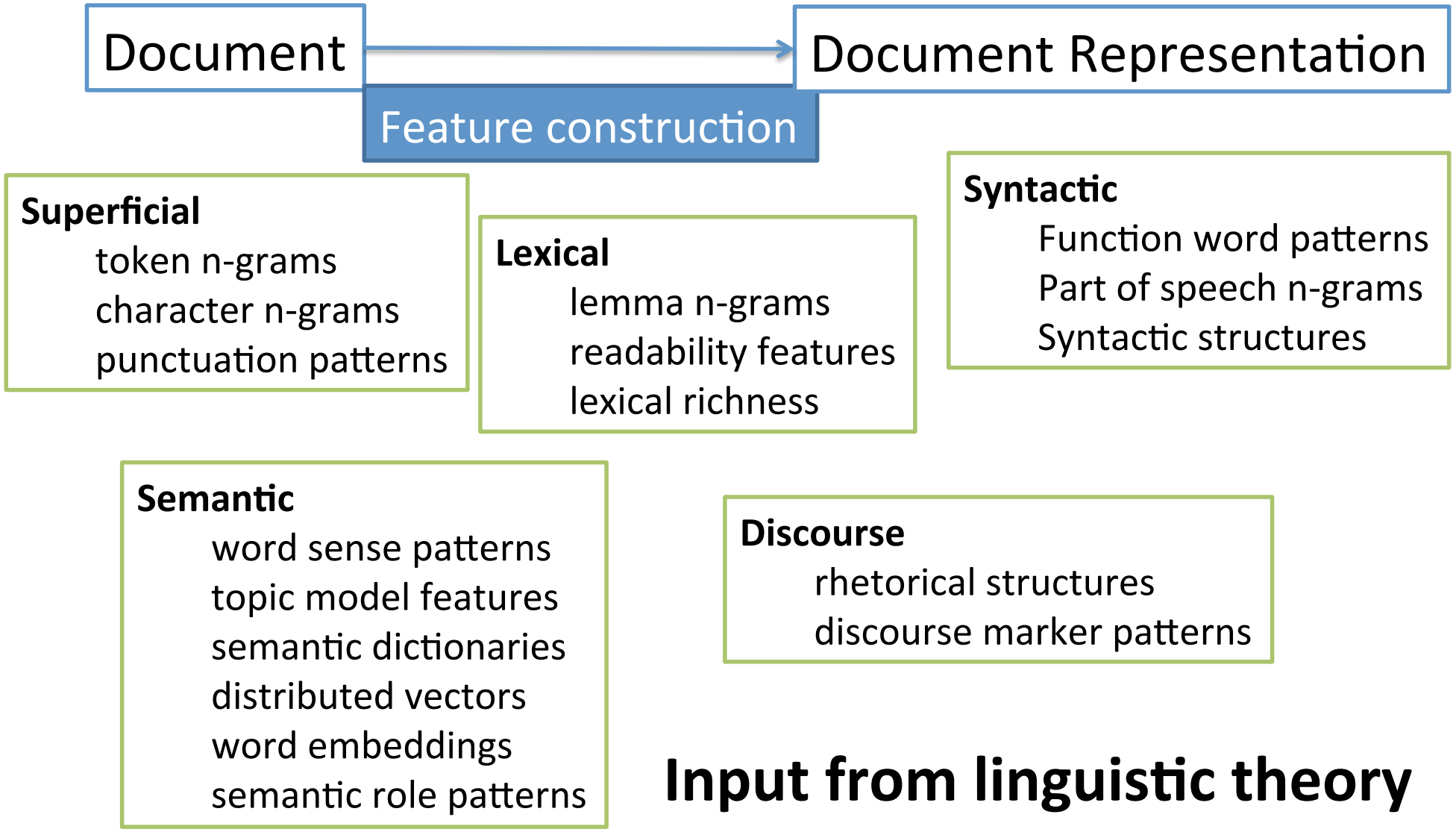
$$p(i \mid l) = \frac{p(l \mid i)p(l)}{p(i)}$$

Bayes rule (1763)

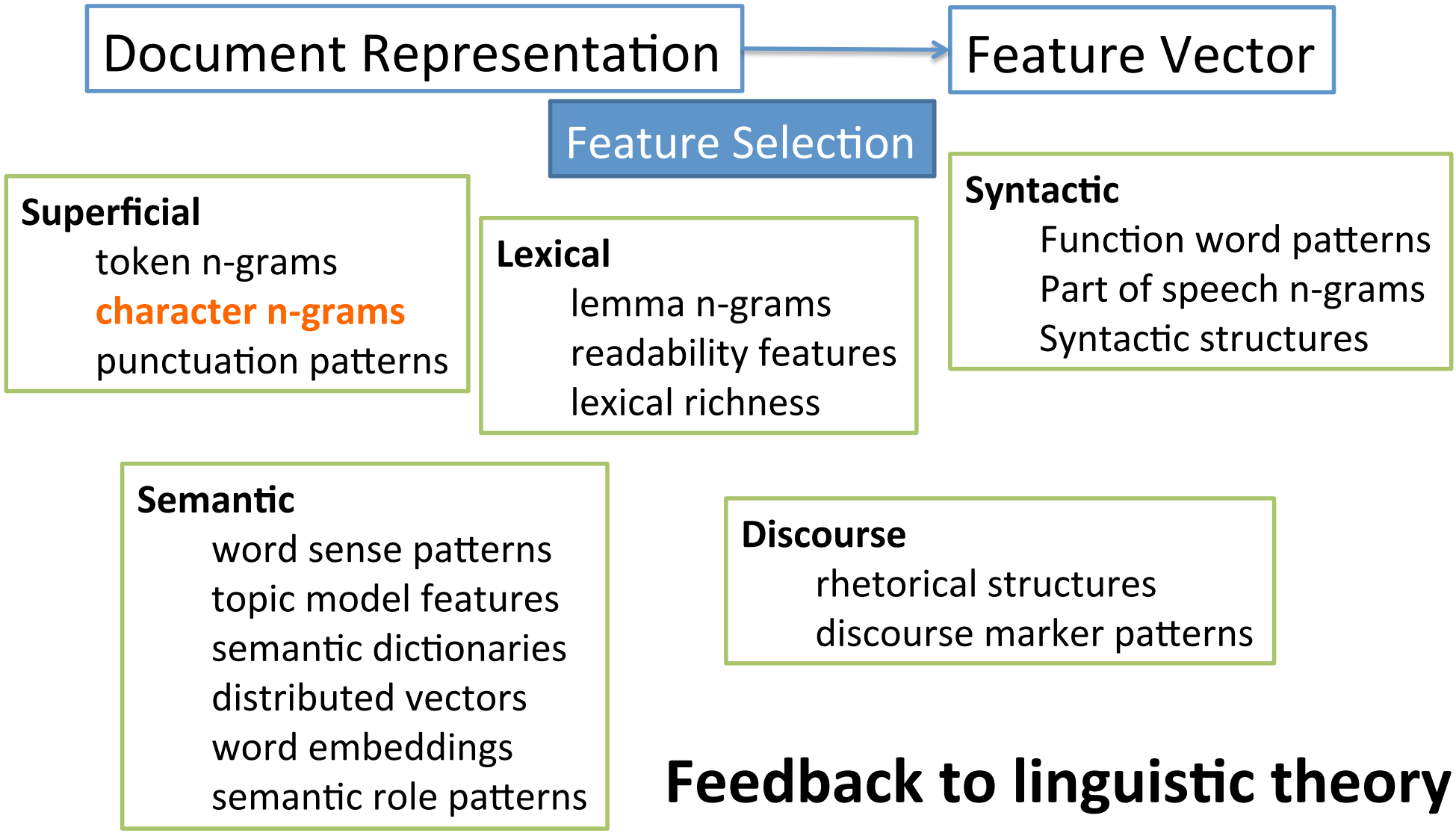
Research in Computational Stylometry

- Basic questions
  - Does an idiolect / stylome exist and (how) can it be measured?
  - Can we handle genre, register & topic interference?
  - Can we handle within-profile interference?
  - Robustness of detection (adversarial stylometry)
  - Dynamic aspects (language change over time with age, illness, language input, …)
- Applications
  - Literary science, forensics / cybersecurity, plagiarism detection, social psychology, sociolinguistics, human resources, marketing, medical diagnostics, …

Method: Advanced Spam Filtering!

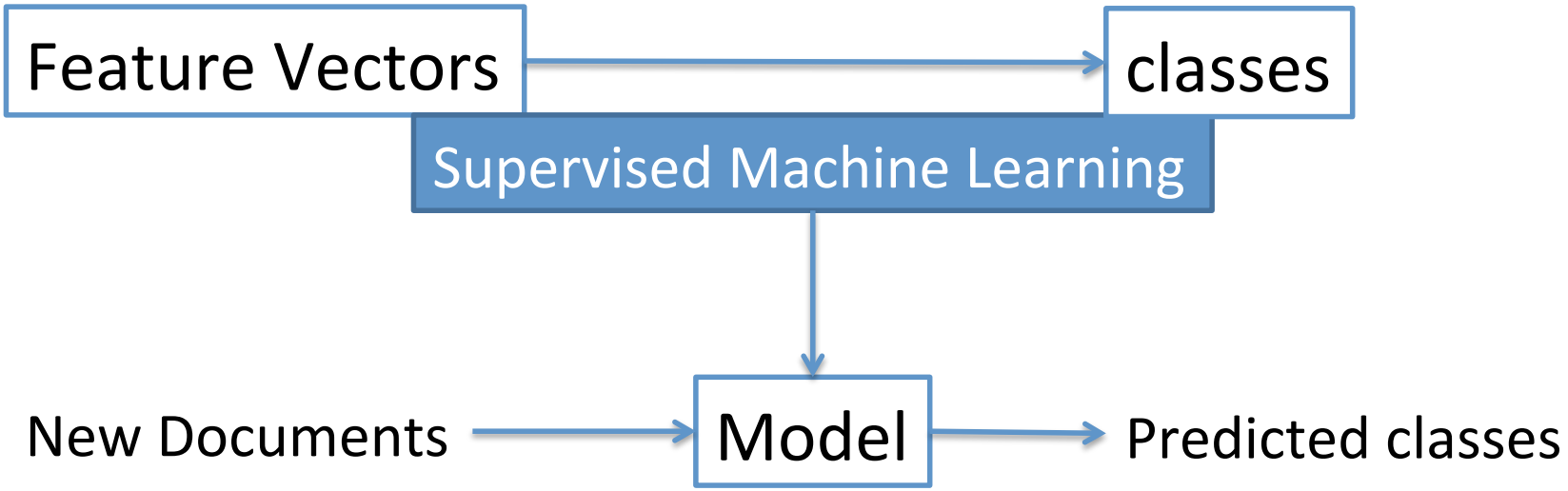


Method: Advanced Spam Filtering!



Feedback to linguistic theory

Method: Advanced Spam Filtering!



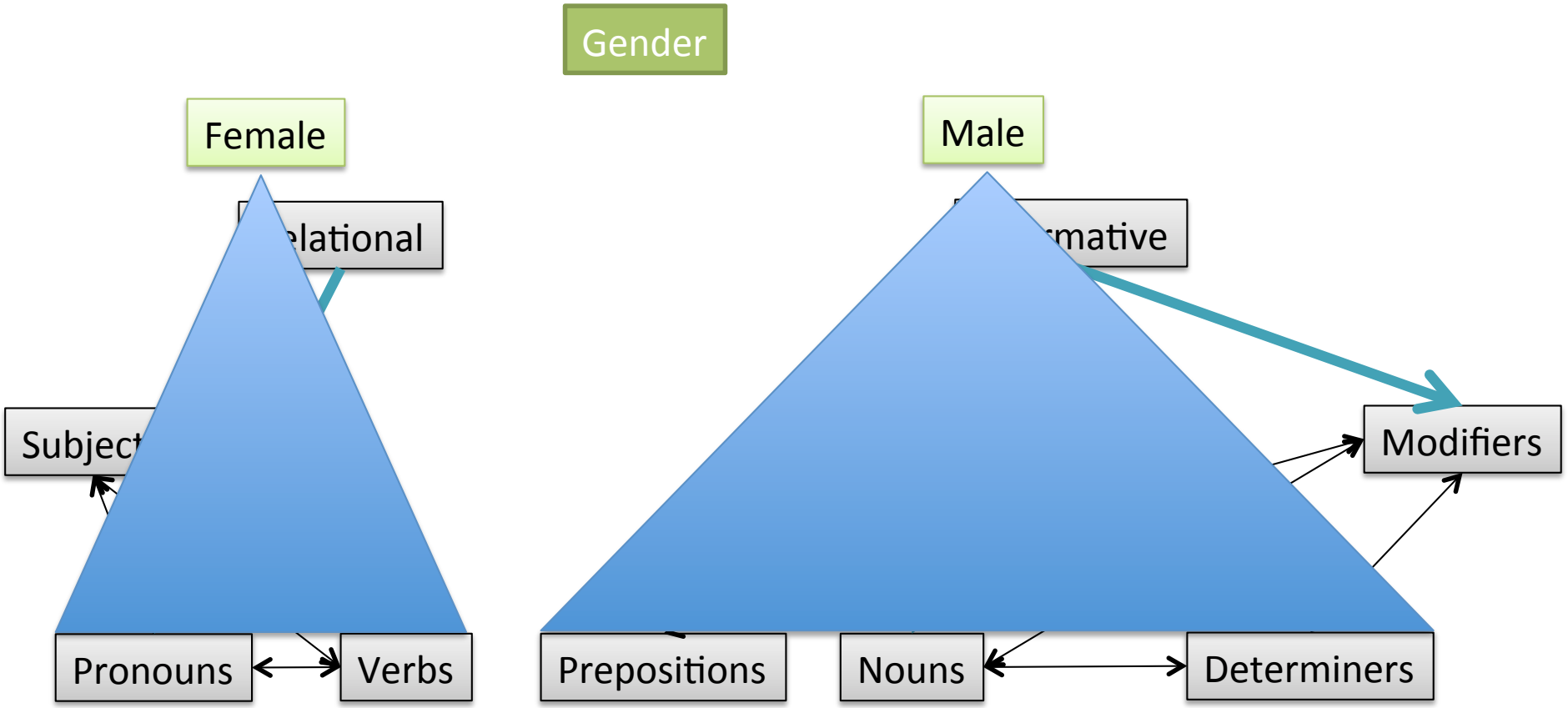
- Objective evaluation on the basis of "gold standard data" and evaluation metrics

Feedback to linguistic theory

Explanation in Stylometry

- **But** quantitative evaluations and lists of important features don't tell us a lot about how it works
- *Why* can age, gender, personality, authorship etc. be determined with a particular set of linguistic features?
- Landmark: gender in fiction and non-fiction
  - Koppel & Argamon et al. 2002


Explanation in Stylometry



Problem

- Same features are informative for different author aspects (e.g. personality & gender)
  - Women ~ extraverts
    - positive and negative emotion words, pronoun use, …
  - Joint learning, ensemble methods, …
- Large **reference corpus** needed of authors with various traits of interest
  - **Sample stratification**
    - Example: CLiPS Stylometry Investigation (CSI) corpus

CLiPS CSI Corpus

- http://www.clips.uantwerpen.be/datasets/csi-corpus
- Genres: essays and reviews (truthful and deceptive)
  - Dutch native speakers, students linguistics and literature
- Meta-data
  - Age, gender, sexual orientation, region, personality
- Yearly expansion
  - Currently: 500 authors; 1250 documents; 600,000 words

www.amicaproject.be

- IWT project coordinated by
  - CLiPS (text mining) with
  - MIOS (sociology)
  - LT3 (text mining)
  - IBCN (software development), and
  - VISICS (image processing)

CLiPS Computational Linguistics & Psycholinguistics University of Antwerp

MIOS Media & ICT in Organisations & Society University of Antwerp

language and translation technology team

IBCN

PSI VISICS

- Combine text analytics, image and video analysis, and data mining

www.amicaproject.be

- Goals
  - Detect situations that are harmful or threatening to young people in social networks
    - Cyberbullying
    - Sexually transgressive behaviour (for example grooming by paedophiles)
    - Depression and suicide announcement
  - Efficient action by moderators, police, parents, peer group, social services, ...
  - Objective measurement, monitoring, trend analysis, …

How urgent is the problem?

- European “Kids online” study (EU, 2011)
  - Age 9-16 in 25 European countries
  - Results
    - Children are 90 minutes per day online
    - Half of them in their bedroom
    - 33% added strangers as friends
    - 15% shared personal information with strangers
      - Including photographs
    - 12% felt they experienced harm

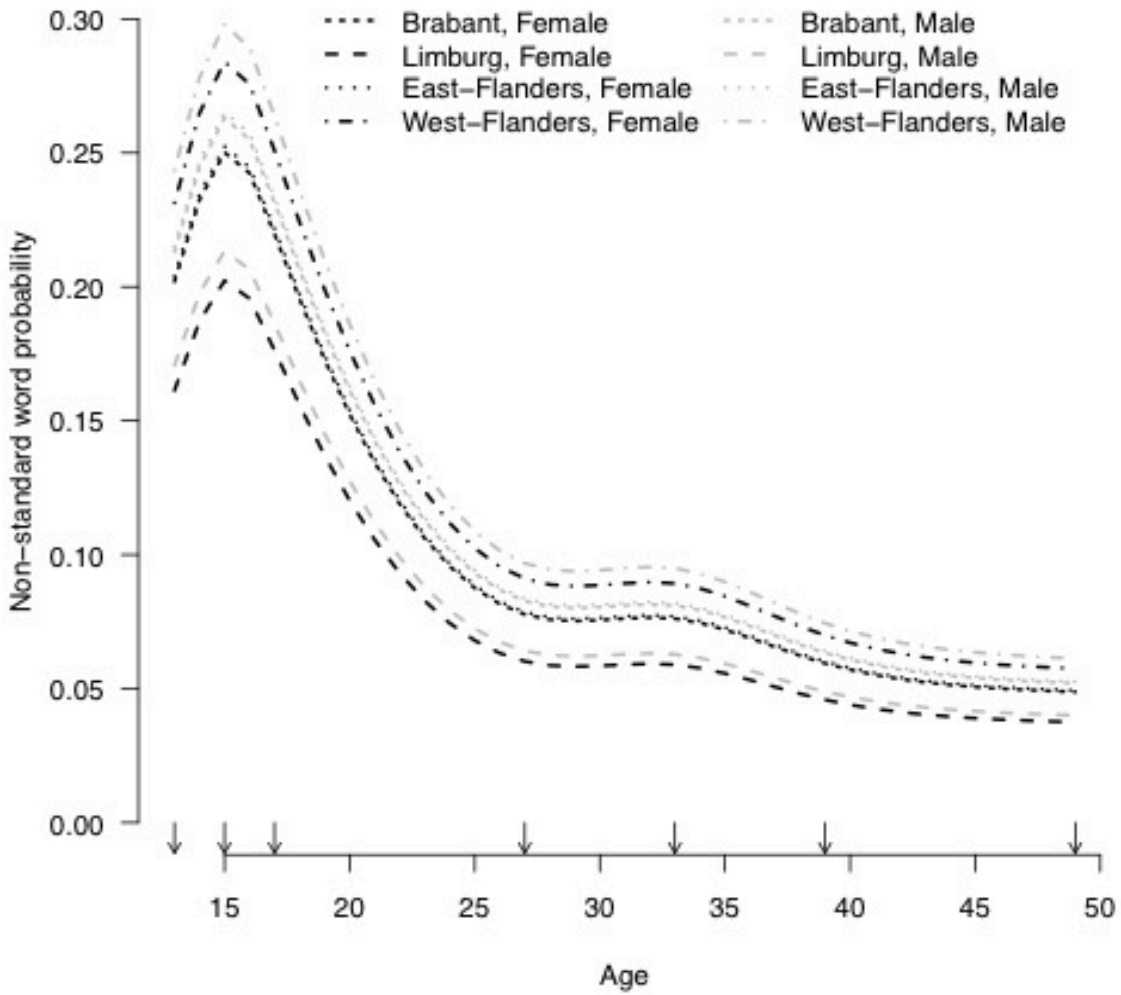
    www.eukidsonline.net

Problem: Properties of social media language

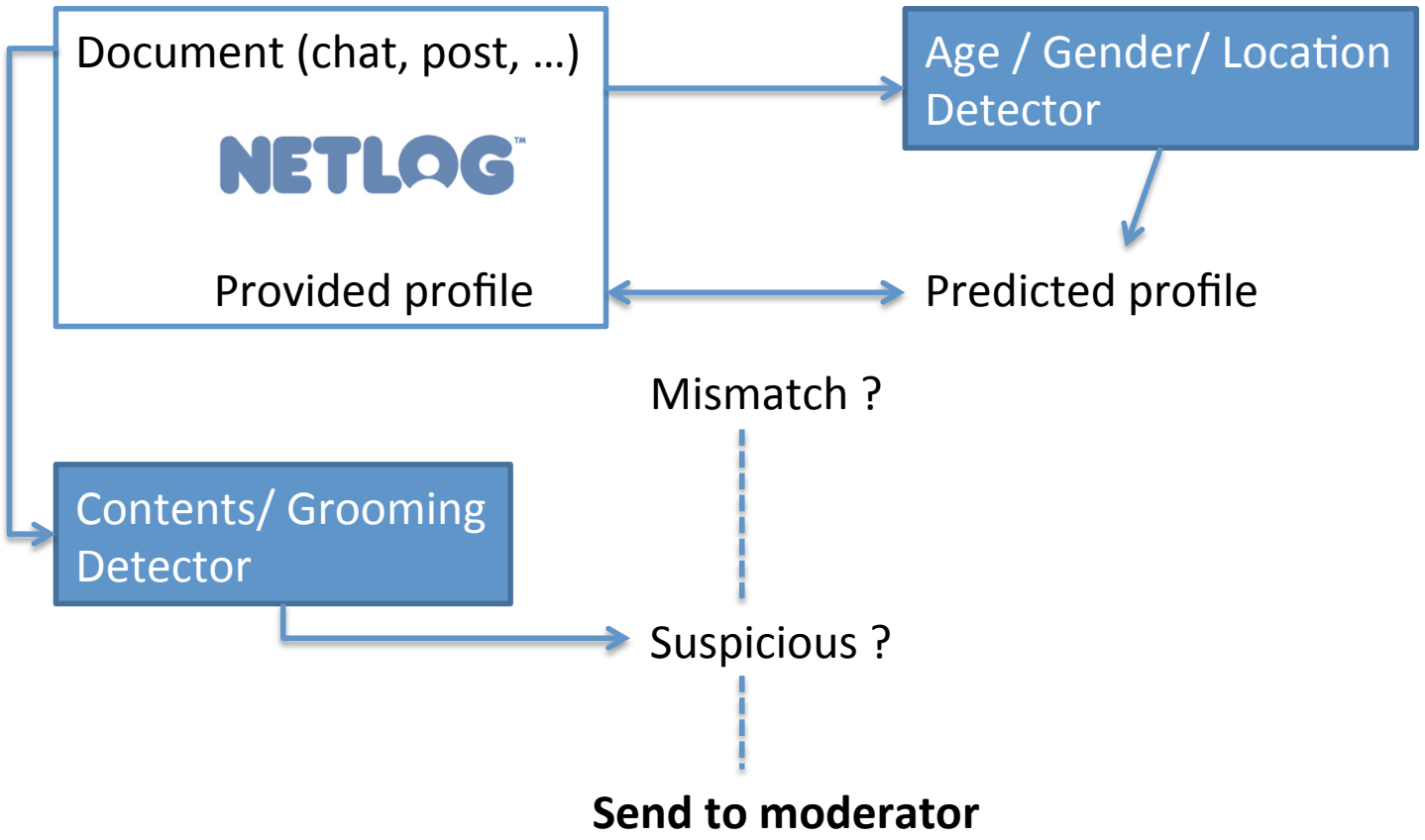
Variation type	Netlog example	Standard Dutch	English
Omission of letters or words	kbda nimr	Ik heb dat niet meer.	I don't have that anymore.
Abbreviations	wrm W8	waarom wacht	why wait
Acronyms	hjg	hou je goed	take care
Character flooding	keiii mooiii	heel mooi	very beautiful
Concatenation	IkKanOokNiiiZonderU!	Ik kan ook niet zonder jou!	I can't live without you either!

+ Dialectical / regional differences

Computational Sociolinguistics?



Pedophile Grooming Detection

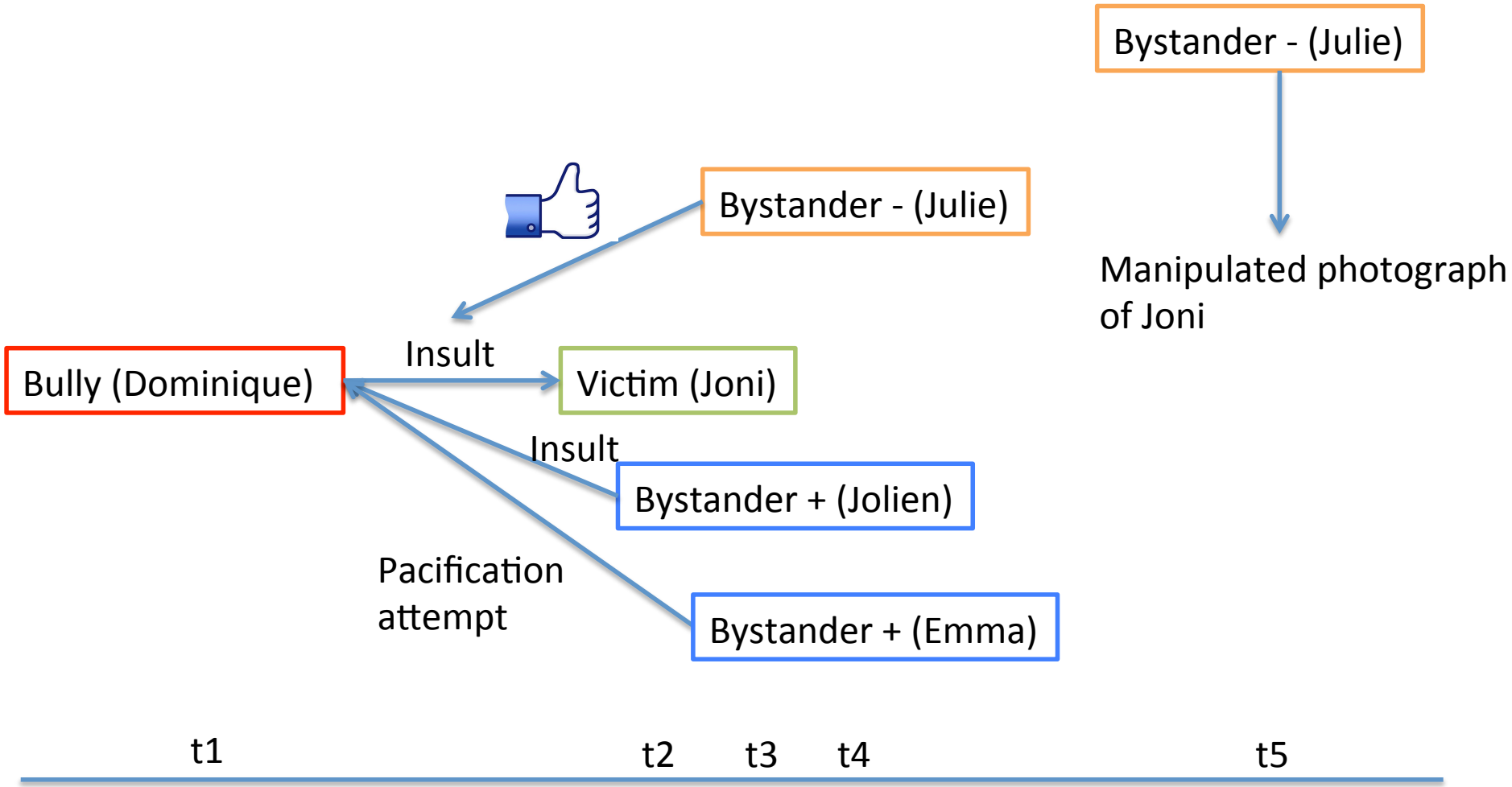

Document (chat, post, …)
NETLOG
Provided profile
Age / Gender/ Location Detector
Predicted profile
Mismatch ?
Contents/ Grooming Detector
Suspicious ?
Send to moderator

Mismatch Classifier

- Netlog data
  - Posts with associated profiles
  - Use profiles to train Age, Gender, Location classifier for mismatch detection
- Results
  - Age ~ 80%, Gender ~ 70%, Location ~ 50%
  - Minus 16 versus plus 25, > 90%
  - Bag of words performs best ☹
    - Different age groups and genders use different intensifiers, emoticons, words in general …

Cyberbullying

Complex Events



Data Collection

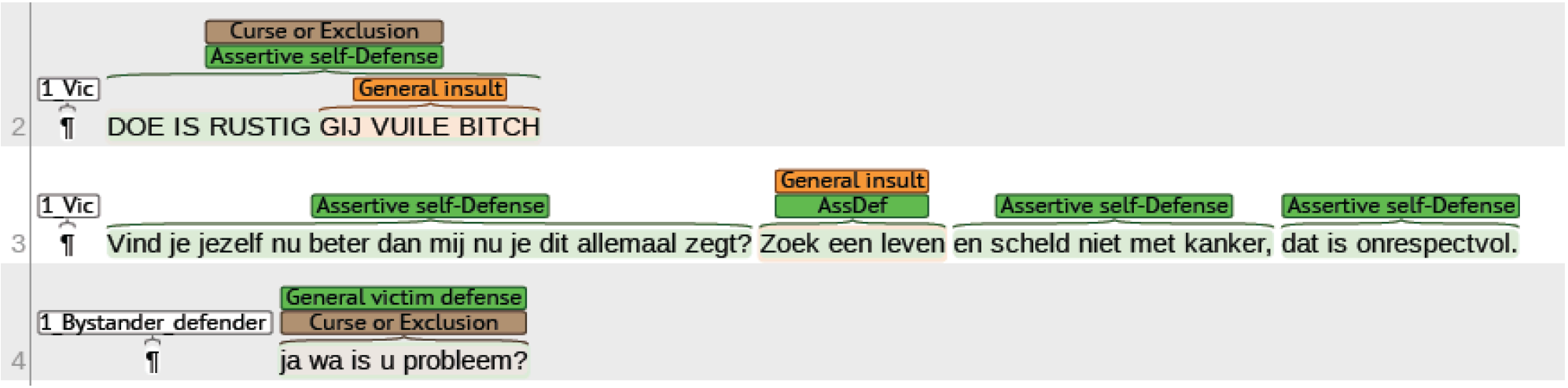
- Media campaign
  - Asking for donation of social network cyberbullying material
  - Massive media response, limited response in data collection
- Role playing in secondary schools (CLiPS + MIOS)
  - Additional goal: education (debriefing)
  - Approach
    - Facebook-like social network
    - Bullying scenarios
    - Profile cards
- Ask.fm

Fine-Grained Annotation (CLiPS + LT3 Ghent University)

- Threat
- Insult
  - Name calling
  - Attacking relatives and friends
  - Discrimination
    - Sexism
    - Racism
  - Curse or Exclusion
  - Defamation
  - Sexual Talk
    - Harmless sexual talk
    - Sexual harassment

Fine-Grained Annotation (CLiPS + LT3 Ghent University)

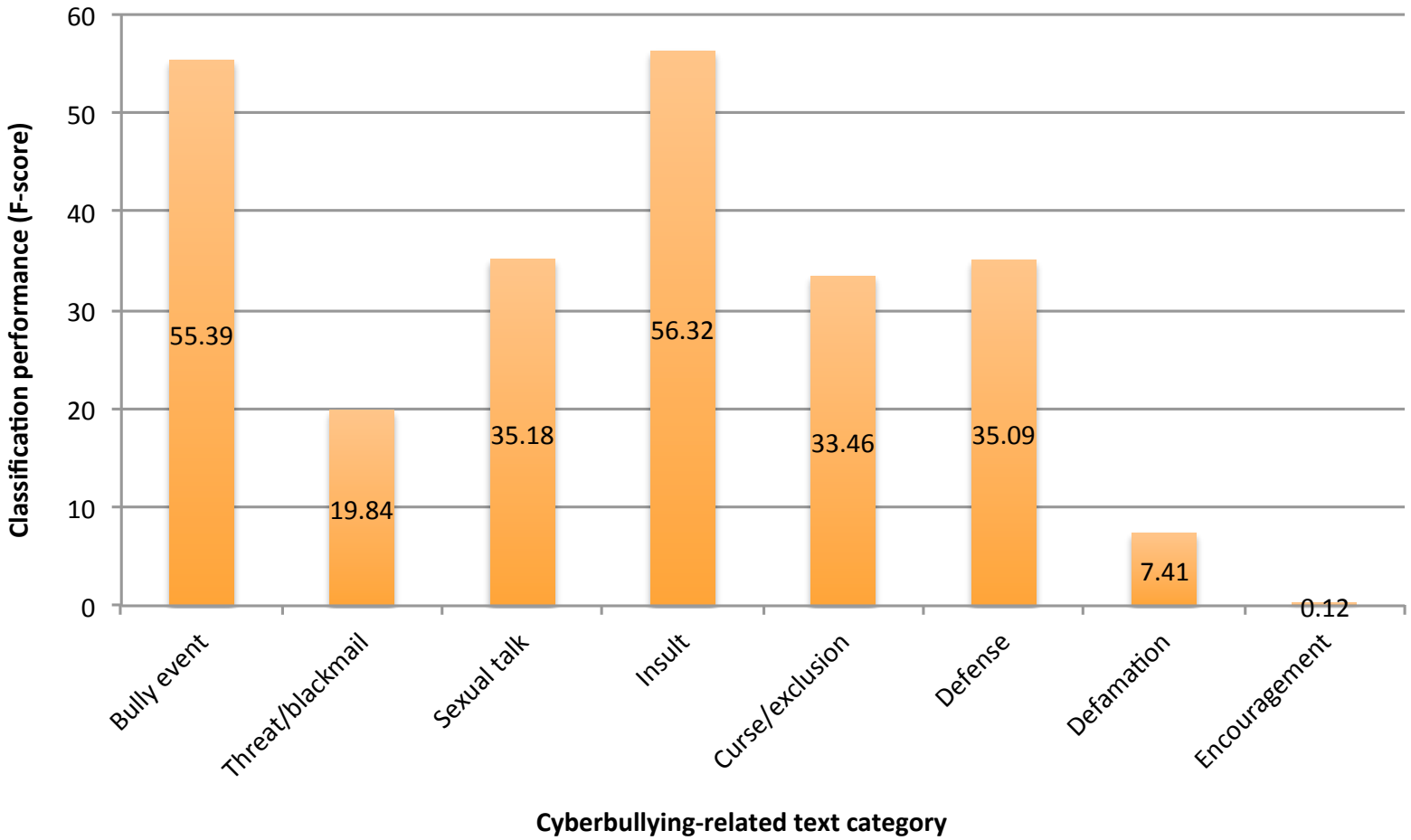
- Defense
  - Bystander defends the victim
    - Good characteristics
  - Victim defends him or herself
    - Assertive
    - Powerless
- Sarcasm
- Other

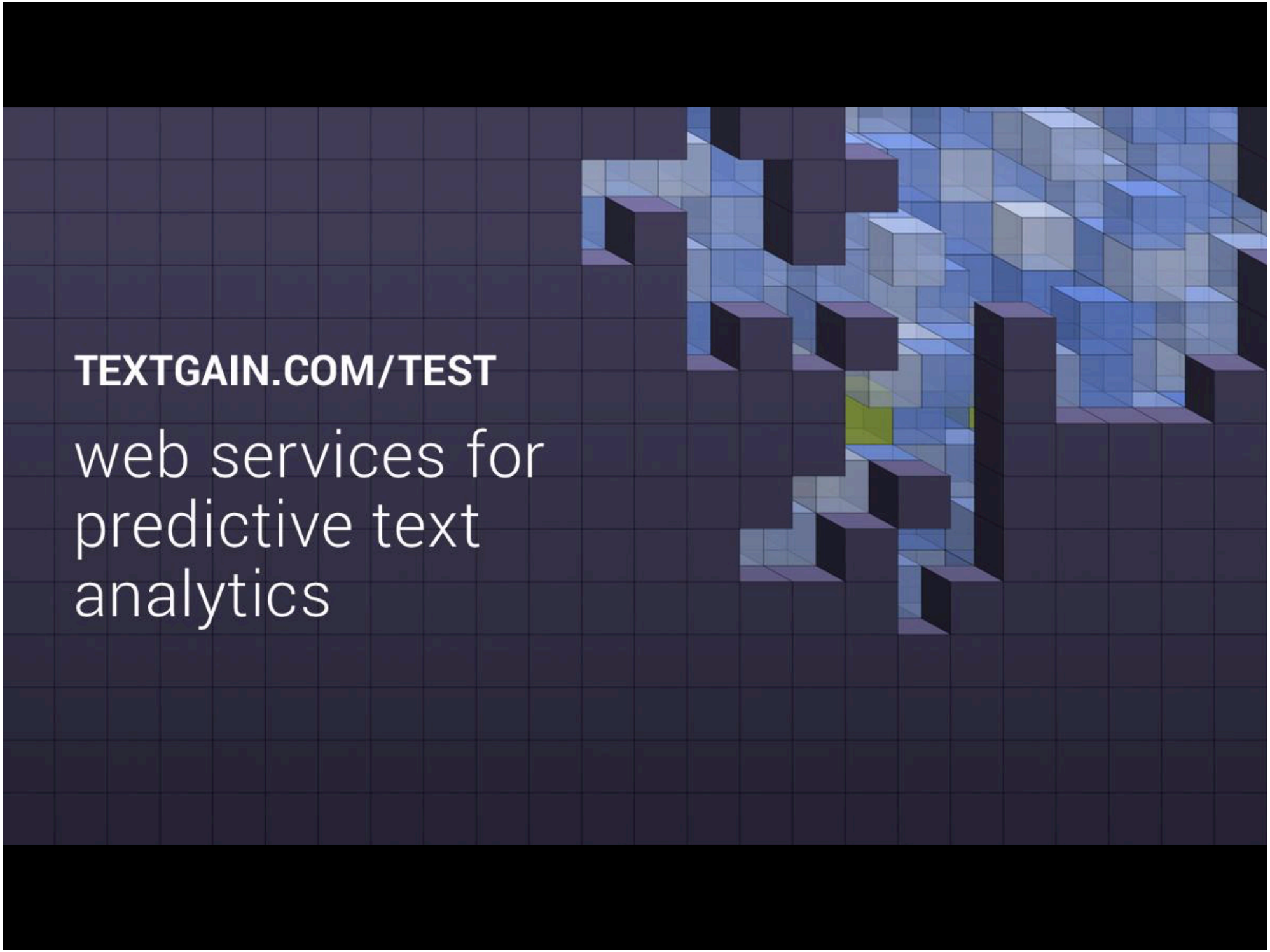


Ask.fm preliminary experiments (Van Hee et al. 2015)

- Predict whether there is bullying and try to predict the different types
- Predictive features used are token and character n-grams and sentiment features
- Simple linear statistical classifier
- ~ 85,000 posts
- Annotation agreement (kappa) 60-65%
- Very skewed data, scarce positive data (~ 10%)
- Best results use all feature types
- Currently, profile features not shown to be helpful

Preliminary Results




TEXTGAIN.COM/TEST
web services for
predictive text
analytics

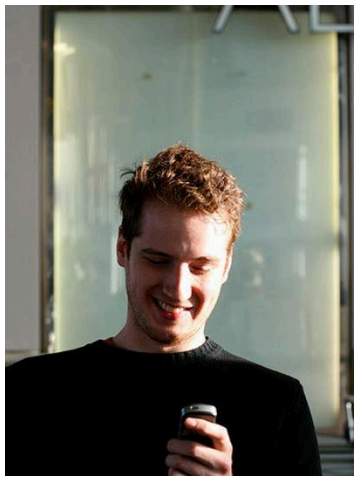
TEXTGAIN

www.clips.uantwerpen.be spin-off

multilingual text analytics tools to extract structured knowledge from unstructured language data

dr. Guy De Pauw
CLiPS senior researcher
language technology
machine learning

dr. Tom De Smedt
CLiPS senior researcher
software engineering (Pattern)
web development



http://textgain.com/test/

UA TechTransfer business developer (support)
Lingua Franca (reseller)

Language identification	available for: af, sq, ar, eu, be, bg, ca, zh, ht, hr, cs, da, nl, en, et, fi, fr, gl, de, el, he, hi, hu, is, id, ga, it, ja, ko, lv, lt, mk, ms, mr, no, fa, pl, pt, ro, ru, sr, sk, sl, es, sw, sv, th, tr, uk, vi, cy, yi												
	en	**nl**	**fr**	**de**	**it**	**es**	**da**	**no**	**sv**	**fi**	**cz**	**pt**	**zh**
Parts-of-speech	✓	✓	✓	✓	✓	✓	□		□	□	□	□	□
Concepts	✓	✓	✓	✓	✓	✓	□		□	□	□	□	□
Sentiment	✓	✓	✓	□	✓	□	□	□	□	□		□	□
Age	✓	✓		✓		✓							
Gender	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Fine-grained sentiment	□	□											
Personality		✓											
Education	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Region		□											

Table 1: overview of current state of development of Textgain technologies.
✓ marks a finished component,
□ marks a component to be deployed in Q2/Q3-2015

APPLICATION AREAS

BRAND MANAGEMENT (demographic product affiliation)

E-MARKETING (social forecast)

DIGITAL PUBLISHING (adaptive content)

CYBERSECURITY, LONGITUDINAL TREND ANALYSIS, E-MAIL ROUTING, RECOMMENDER SYSTEMS, ONLINE DATING, ...

Conclusions

- Profiling is a Computational Stylometry application with interesting research questions and useful applications (AMiCA)
- And possibly some commercial value already (TEXTGAIN)