Personality traits on Twitter for less-resourced languages

Barbara Plank[1], Ben Verhoeven[2] & Walter Daelemans[2]

[1] CST, University of Copenhagen, Denmark

[2] CLiPS Research Center, University of Antwerp, Belgium

Presented at CLIN 26, Amsterdam
December 18, 2015

Personality

- “Individual differences between people with respect to patterns of behavior, cognition, and emotion” (Michel, Shoda & Smith, 2004)
- Described in scaled components
- Different typologies
  - Big Five (OCEAN)
  - Myers-Briggs Type Indicator (MBTI)

Personality

- Big Five
  - Openness to experience
    - Inventive/curious vs. consistent/cautious
  - Conscientiousness
    - Efficient/organized vs. easy-going/careless
  - Extraversion
    - Outgoing/energetic vs. solitary/reserved
  - Agreeableness
    - Friendly/compassionate vs. analytical/detached
  - Neuroticism (emotional stability)
    - Sensitive/nervous vs. secure/confident

Personality

- MBTI
  - Extraversion vs. Introversion
  - iNtuitive vs. Sensing
  - Thinking vs. Feeling
  - Judging vs. Perceiving

- 16 Types
  - E.g. ESTJ, ISFP, ENTP, …

Existing resources

Corpus	Authors	Year	Language	Size	Open
Essays	Pennebaker & King	1999	EN	2,479 docs	x
myPersonality	Kosinski & Stillwell	2007	EN	millions	
Personae*	Luyckx & Daelemans	2008	NL	145 docs	x
Blogs	Iacobelli et al.	2011	EN	3000 authors	
WCPR13	Celli et al.	2013	EN	10,000 posts	x
YouTube Vlogs	Biel & Gatica-Perez	2013	EN	404 docs	x
PAN 2015	Rangel et al.	2015	EN, ES, NL, IT	±500 authors	x

* Only one with MBTI types, others use Big Five

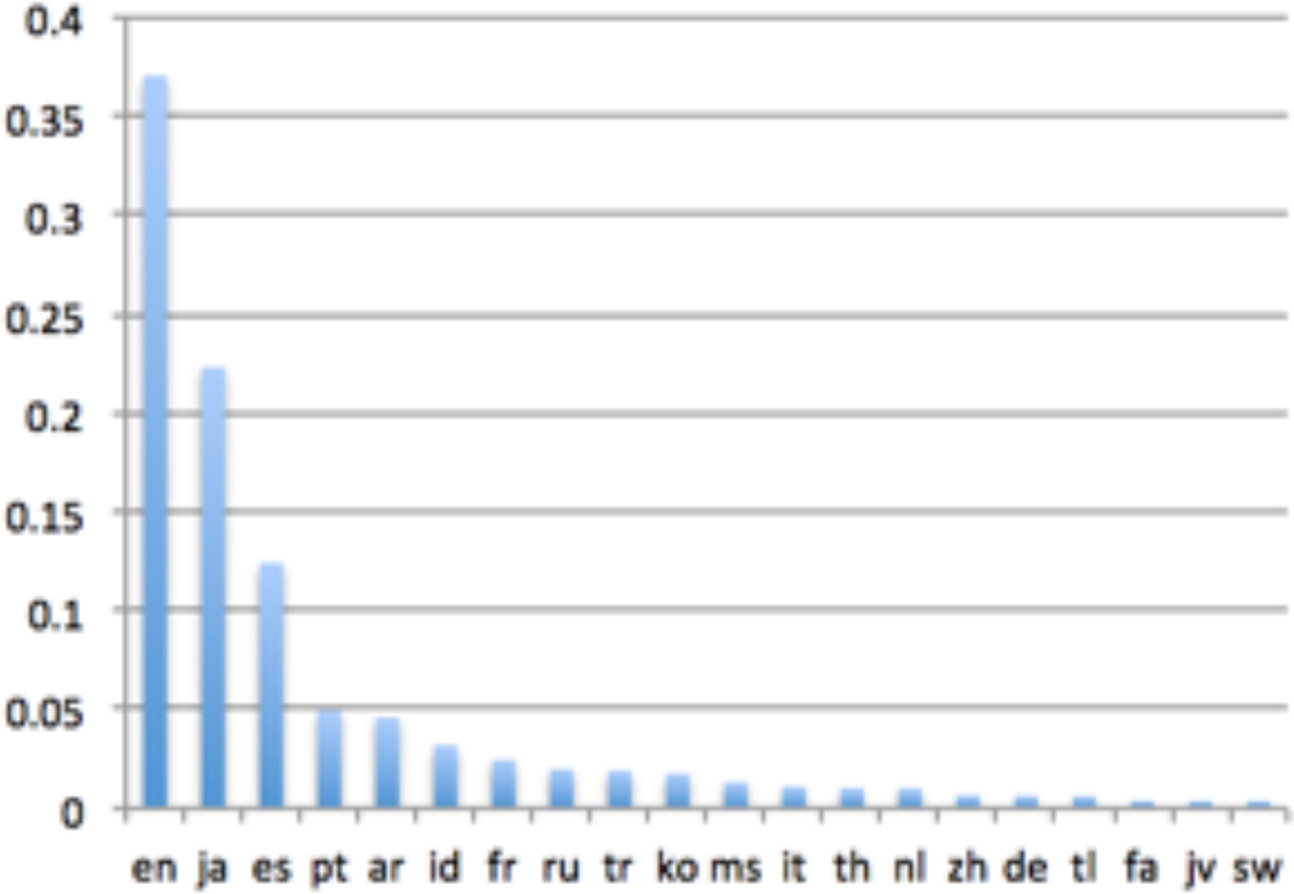
Building on previous efforts

- CLiPS Stylometry Investigation (CSI) corpus
  - Verhoeven & Daelemans (2014)
  - Continuous effort
  - Over 1,200 reviews and 500 essays/papers
  - Rich metadata
    - Big Five & MBTI

Building on previous efforts

- Plank & Hovy (2015)
  - Twitter mining for only one week
  - Searching for MBTI types via API
  - Only English
  - Annotating gender
  - Result
    - 1500 authors
    - 1.2M tweets

Twitter language distribution



Approximated with `langid` (Lui & Baldwin, 2012) on a Twitter sample of 65m tokens

Less-resourced languages

- ≠ low-resourced languages
- Italian, Dutch

- Can we use the Plank & Hovy (2015) approach to do large-scale personality detection on languages that are less present on Twitter?

Yes!

Data collection

- Twitter search instead of mining through API
- Search for combination of each MBTI type with language-specific words
  - – IT: *che, fatto, sono*
  - – NL: *ik, jij, het, persoonlijkheid*
- Download HTML

Data Clean-Up

- Filter out tweets that were not relevant:
  - Not about author
    - *@schrooten ok, ik heb deze test destijds met een uitgebreide vragenlijst op mijn werk gedaan. Meerdere van mijn collega PM-ers zijn ESTJ...*
  - Ambiguity of type
    - *Volgens mij ben ik zowel INTJ als ESTJ -- het eerste als ik me rot voel, het tweede als het goed gaat. #beetjevreemd*
  - In different language
    - ***Estj** seregas muzon4**ik**? **Het**. O, nu tad davaj daj timati, etoj dj dljee.;D*
- Label for gender

Some Statistics

	Profiles	Tweets	Tokens
Italian	370	700 K	8.8 M
Dutch	577	1.2 M	13 M

- IT: biased to female introvert
- NL: gender balance but extravert

Experiments on user level

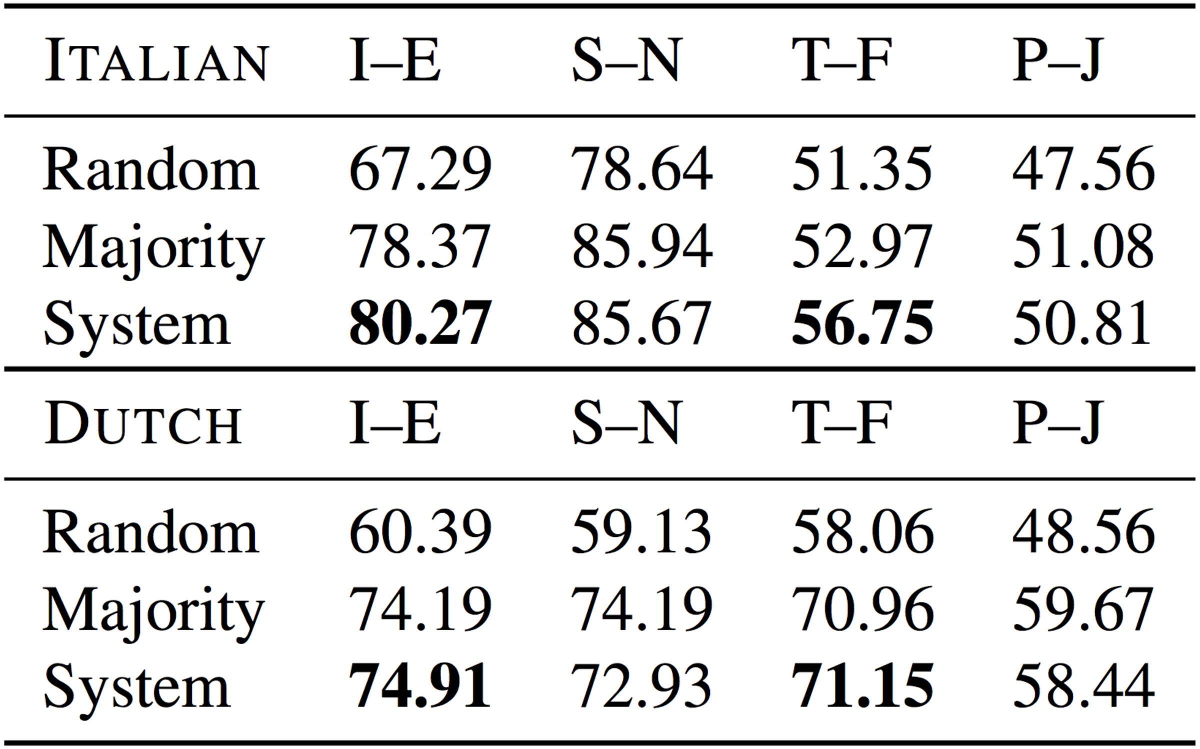
- **Instances**: concatenation of tweets
- **Model**: Logistic Regression with SKLearn
- **Preprocessing**:
  - Tokenization
  - Replacement of URLs, hashtags and mentions by unique token respectively
  - Remove tweets with MBTI type
- **Evaluation**: tenfold cross-validation

Experiments on user level

- **Features**:
  - Word n-grams
  - Character n-grams
  - Counts of Twitter profile
    - Tweets
    - Followers
    - Statuses
    - Favorites
    - Listed

Experiments on user level

- **Results**



ITALIAN	I–E	S–N	T–F	P–J
Random	67.29	78.64	51.35	47.56
Majority	78.37	85.94	52.97	51.08
System	**80.27**	85.67	**56.75**	50.81
DUTCH	I–E	S–N	T–F	P–J
Random	60.39	59.13	58.06	48.56
Majority	74.19	74.19	70.96	59.67
System	**74.91**	72.93	**71.15**	58.44

Accuracy for four discrimination tasks with up to 2000 tweets/user

Ongoing work

TwiSty Corpus

Twitter Stylometry Corpus for Western European Languages

- Large-scale multilingual corpus for personality and gender
- Open source
- All Western European languages in top 20 of Twitter frequencies, apart from English
    - IT, NL, DE, ES, PT, FR

Context words

Italian	*che, sono, fatto*
Dutch	*ik, jij, het, persoonlijkheid*
German	*ich, bist, Persönlickheit, dass*
French	*suis, c’est, personnalité*
Spanish	*soy, tengo, personalidad*
Portuguese	*sou, personalidade*

Frequent misspellings

In many languages

- INFP = info

In some languages

- ESTP = esto

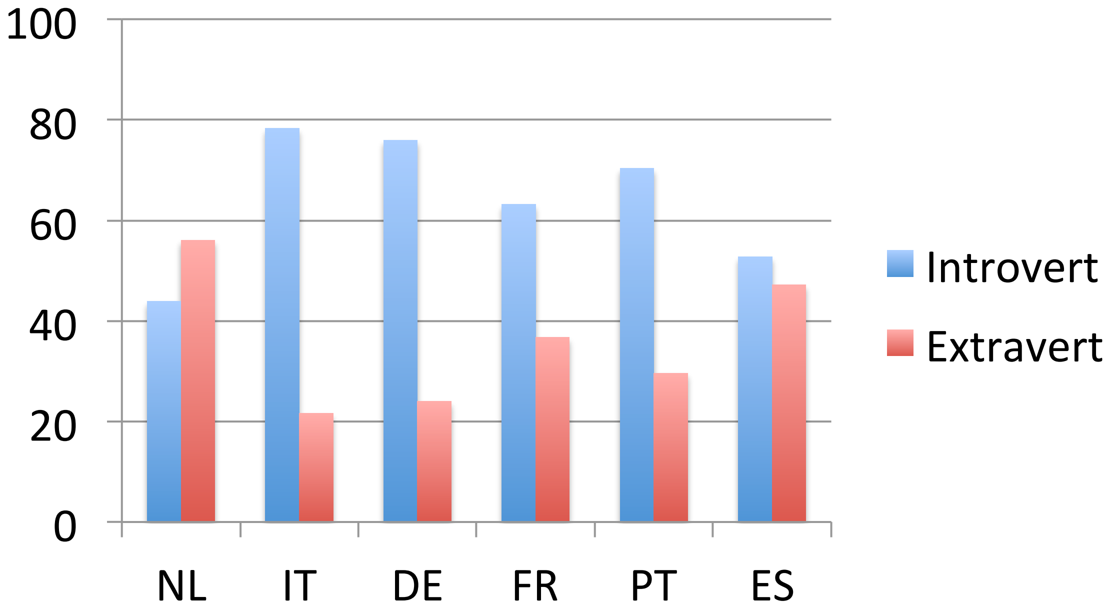
Corpus Statistics

Language	Before Clean-Up	After Clean-Up	# Tweets
Italian*		370	700 K
German	1,457	411	950 K
Dutch	2,691	1,000	2 M
French	4,982	1,417	-
Portuguese	12,914	4,375	-
Spanish	21,731	-	-

*To be redone with same methodology

Introversion vs. Extraversion

- More introverts than extraverts for all languages, except Dutch
  - Any ideas?

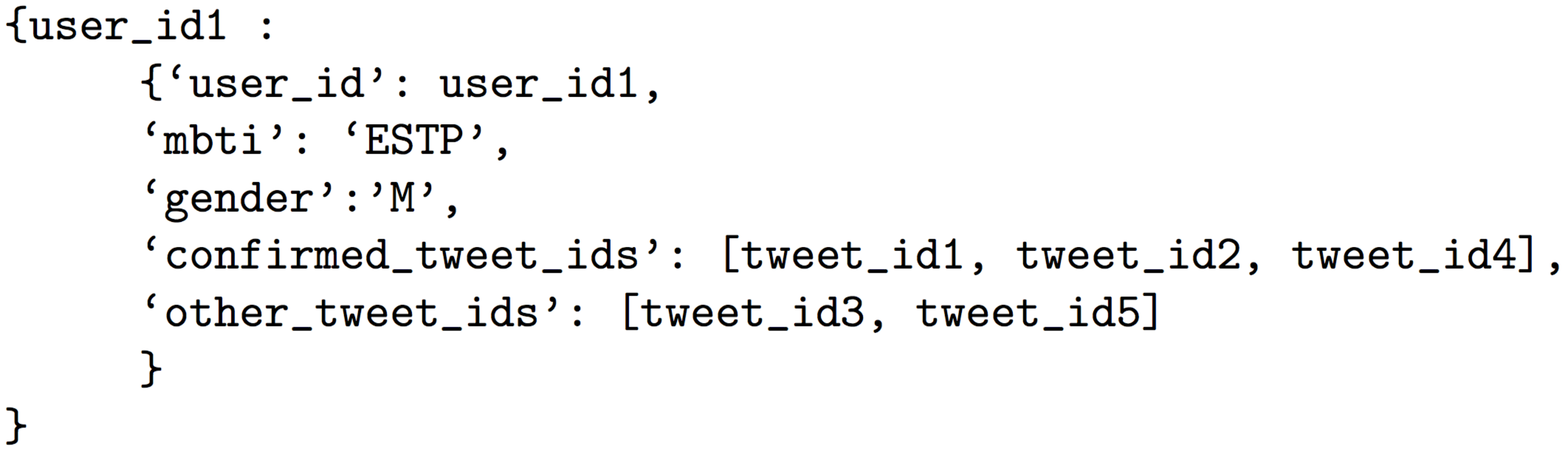


Language Identification

- Many bilingual/polyglot Twitter users
- Tweet-level identification
- Majority voting approach with three language identifiers
- Dutch and German: ± 74%

Tool	Authors	# Langs
langid.py	Lui & Baldwin (2012)	97
langdetect	Nakatani (2010)	53
ldig	Nakatani (2012)	17

Corpus Structure



```
{user_id1 :
      {‘user_id’: user_id1,
      ‘mbti’: ‘ESTP’,
      ‘gender’:’M’,
      ‘confirmed_tweet_ids’: [tweet_id1, tweet_id2, tweet_id4],
      ‘other_tweet_ids’: [tweet_id3, tweet_id5]
      }
}
```

Questions?

- Thanks for your attention.