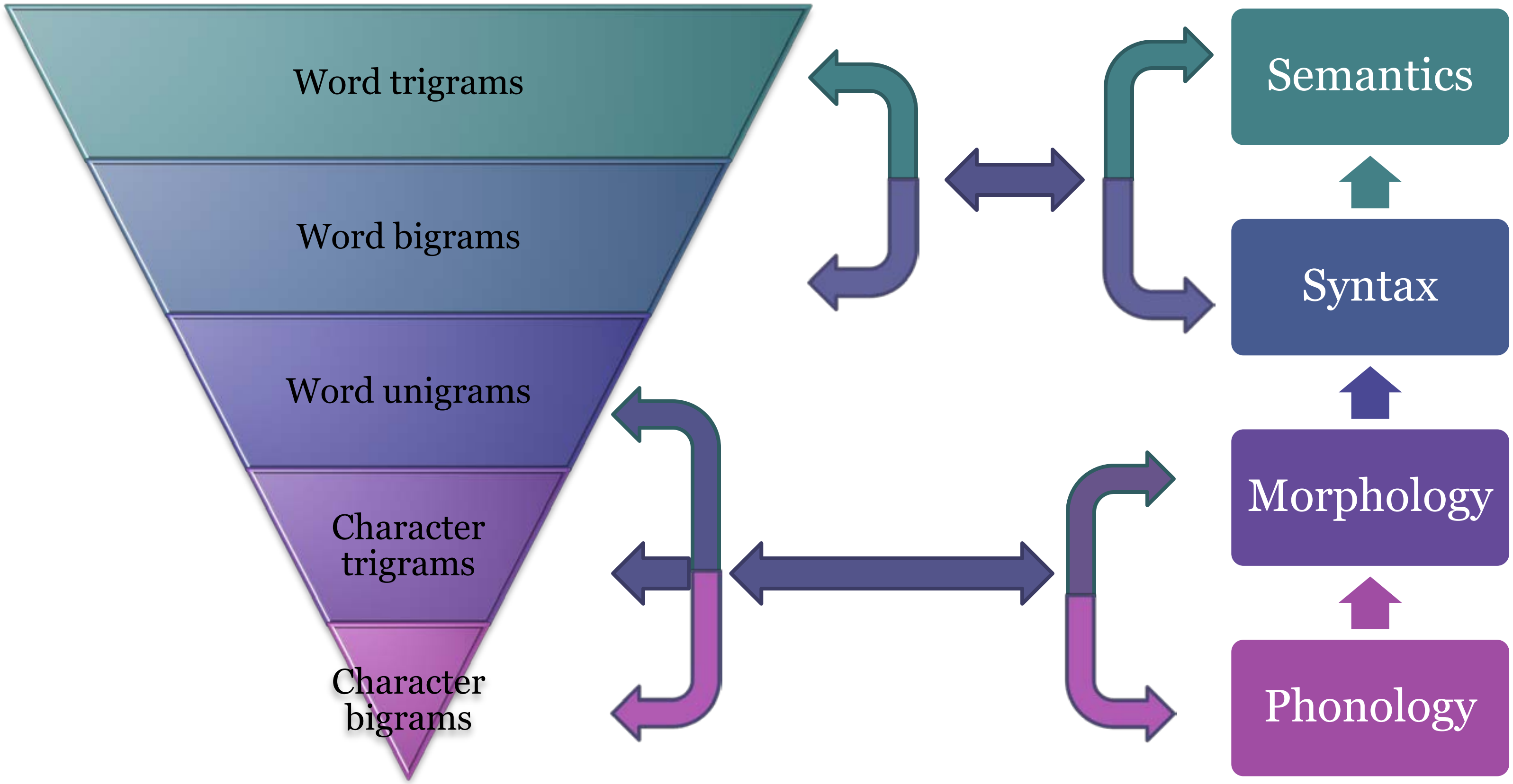
Authorship identification in large email collections: Experiments using features that belong to different linguistic levels

George K. Mikros & Kostas Perifanos
National and Kapodistrian
University of Athens

Style

- Our approach to authorship identification is based mainly on the idea that an author's style is a complex multifaceted phenomenon affecting the whole spectrum of his/her linguistic production.
- Following the old theoretical notion of "double articulation" of the Prague School of Linguistics we accept that stylistic information is constructed in parallel blocks of increasing semantic load, from character n-grams, to word n-grams.
- In order to capture the multilevel manifestation of stylistic traits we should detect these features, which belong to many different linguistic levels, and utterly combine them for achieving the most accurate representation of an author's style.

An hierarchical representation of features and related linguistic levels



Features

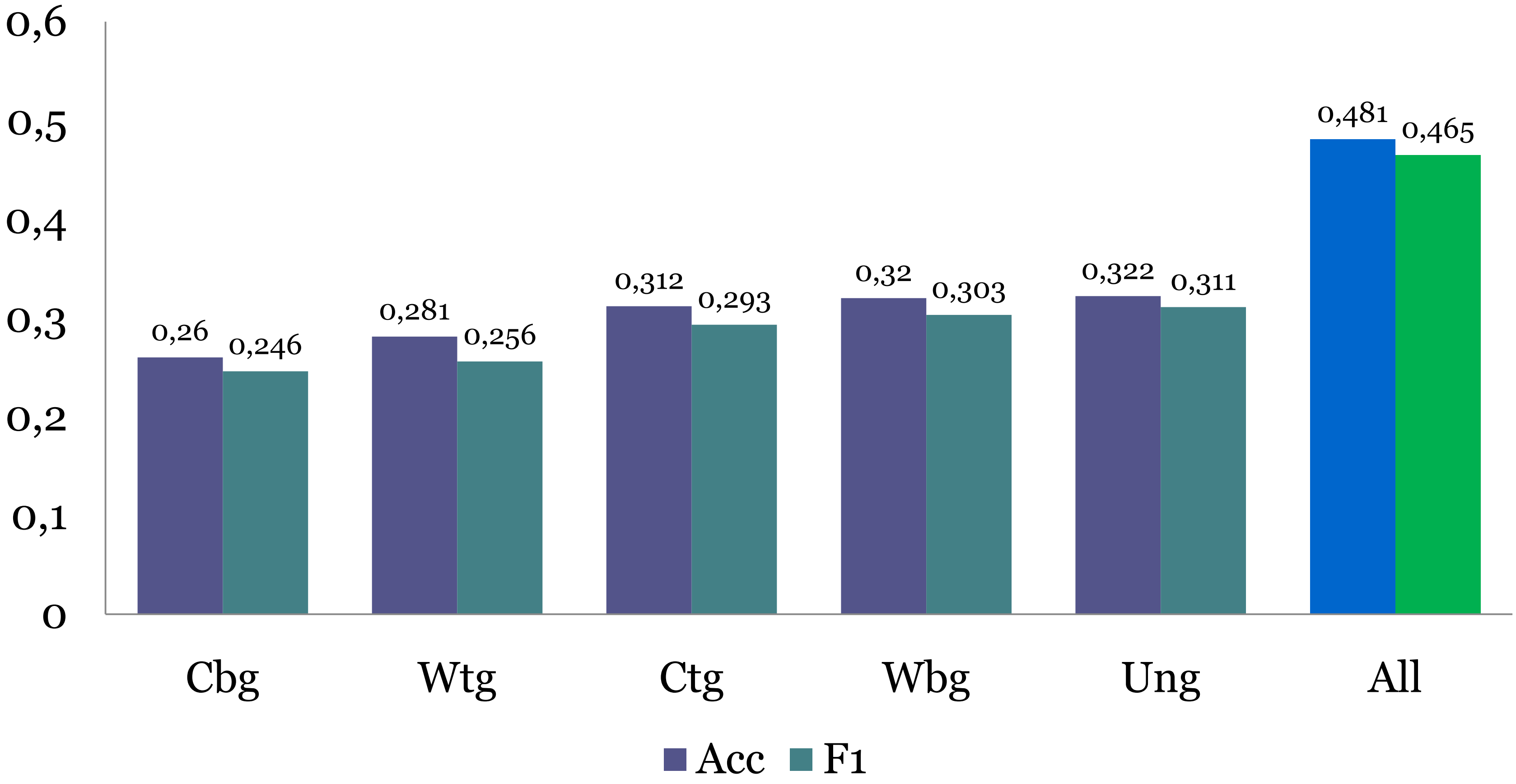
1000 most frequent n-grams from the following feature groups:

- **Character Bigrams (cbg)**: Character n-grams provide a robust indicator of authorship and many studies have confirmed their superiority in large datasets.
- **Character Trigrams (ctg)**: Character trigrams capture significant amount of stylistic information and have the additional merit that they also represent common email acronyms like FYI, FAQ, BTW, etc.
- **Word Unigrams (ung)**: Word frequency is considered among the oldest and most reliable indicators of authorship outperforming sometimes even the n-gram features.
- **Word Bigrams (wbg)**: Word bigrams have long been used in authorship attribution with success.
- **Word Trigrams (wtg)**: Word trigrams have also been found to convey useful stylistic information since they approach more closely the syntactic structure of the document.

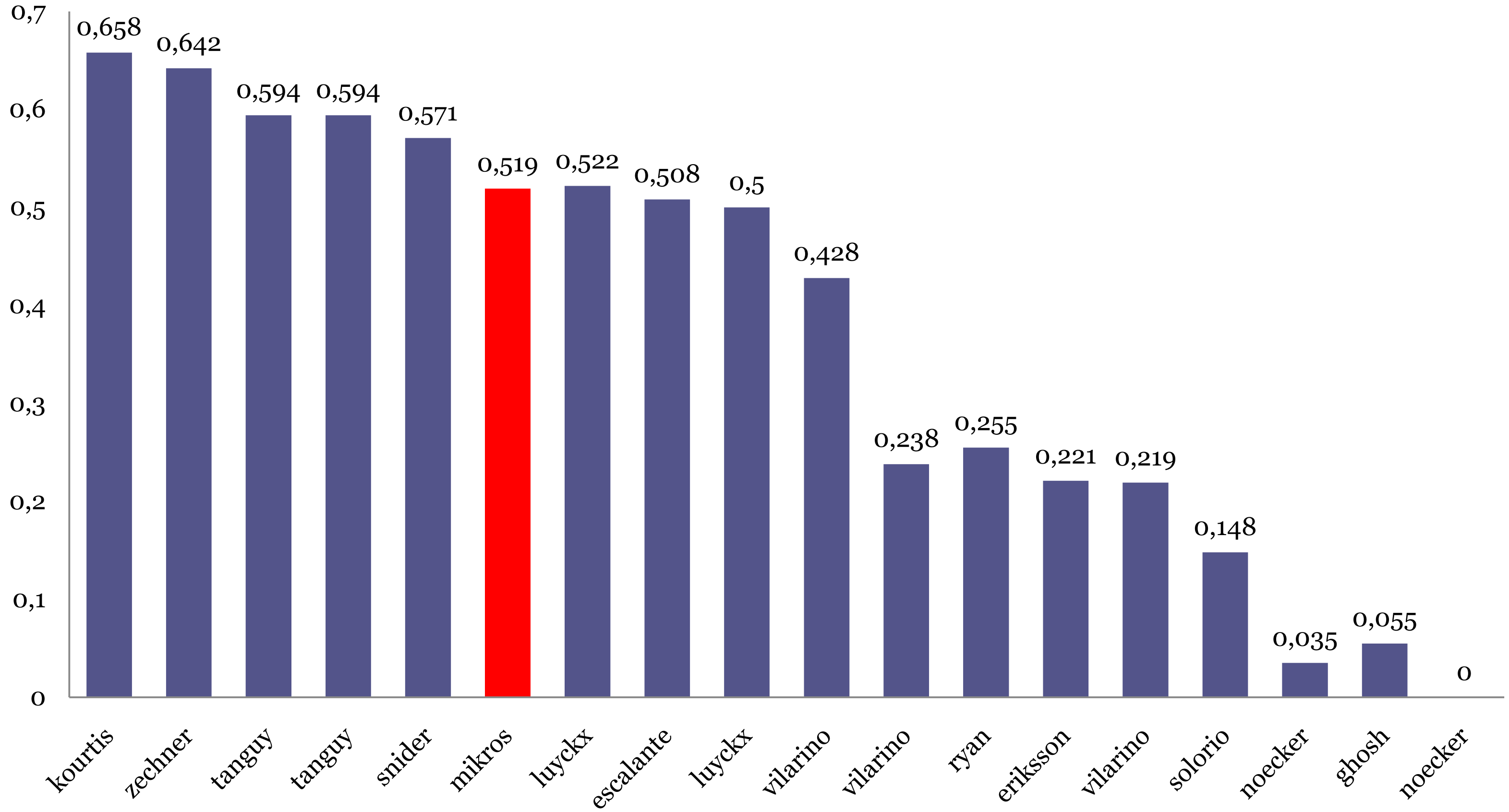
Algorithms and Datasets

- Large and Small Datasets (Authorship Attribution scenario)
  - L2 Regularized Logistic Regression (Authorship Attribution tasks)
- Large and Small + Datasets (Combined Authorship Attribution and Verification scenario)
  - One-Class SVM and L2 Regularized Logistic Regression
- Verify 1, 2 & 3 Datasets (Pure Author Verification)
  - One-Class SVM (Authorship Verification tasks) using only the 2000 most frequent character bigrams.

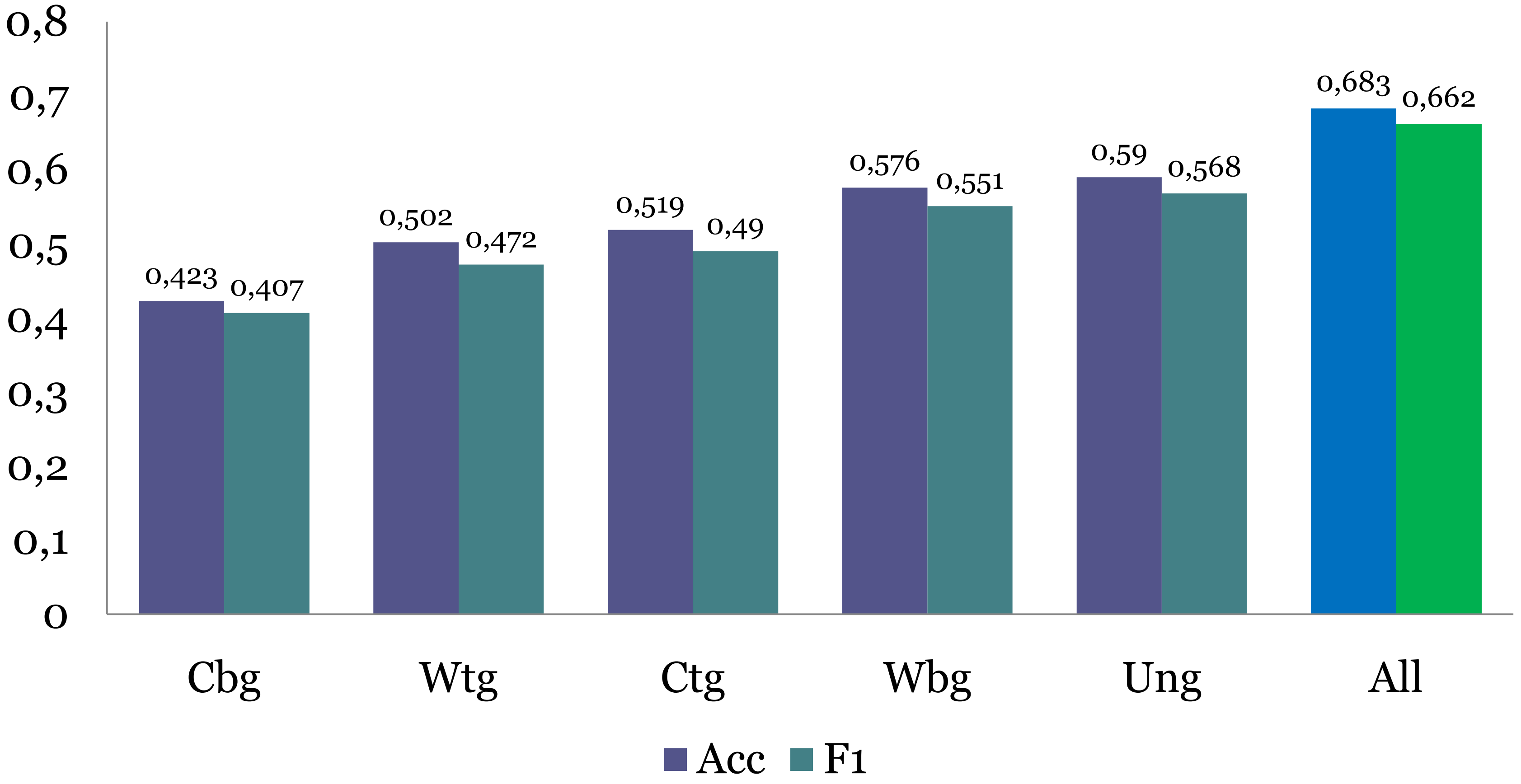
Results in Large Train Dataset



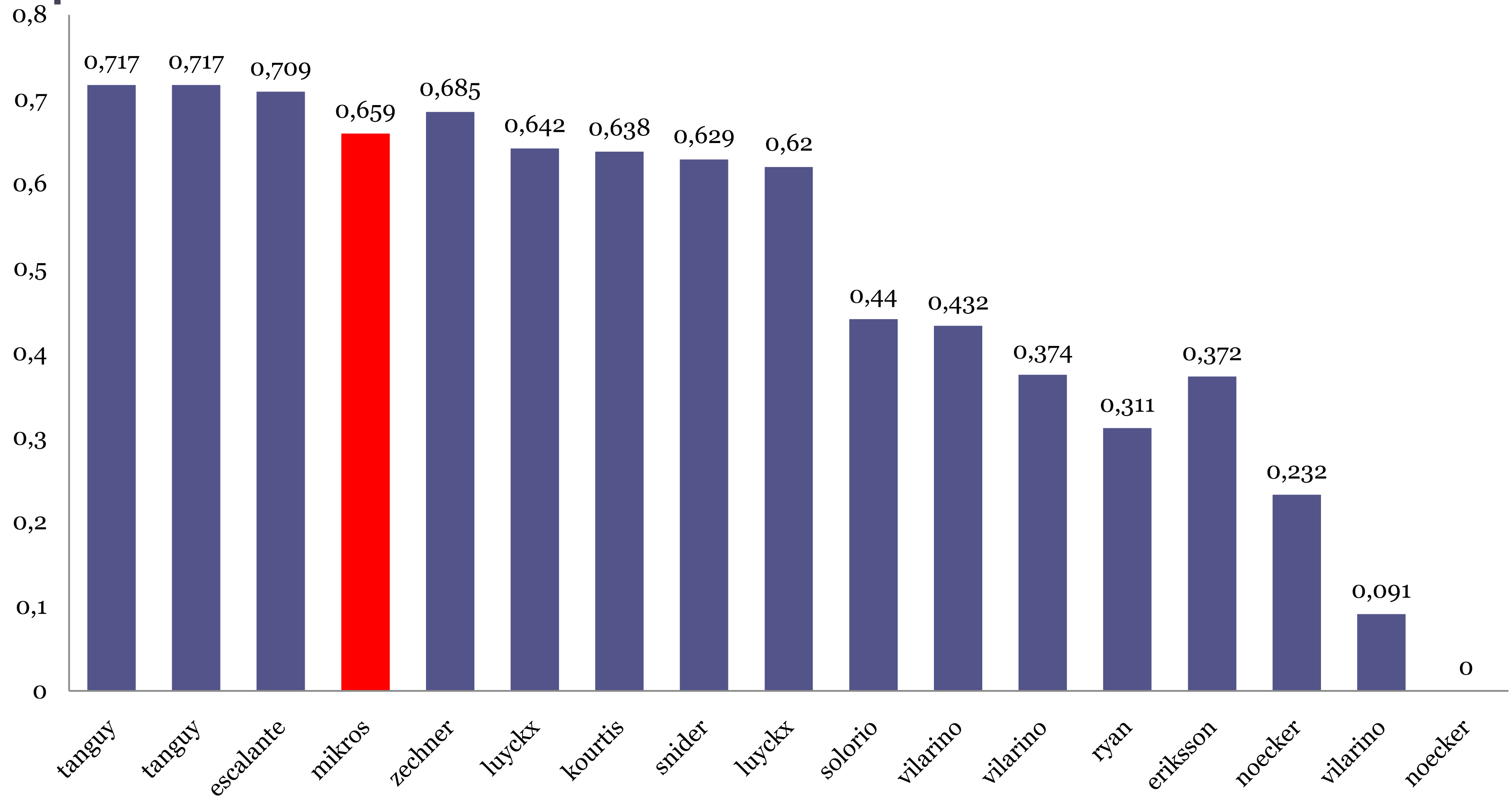
F_1 in Large Test Dataset



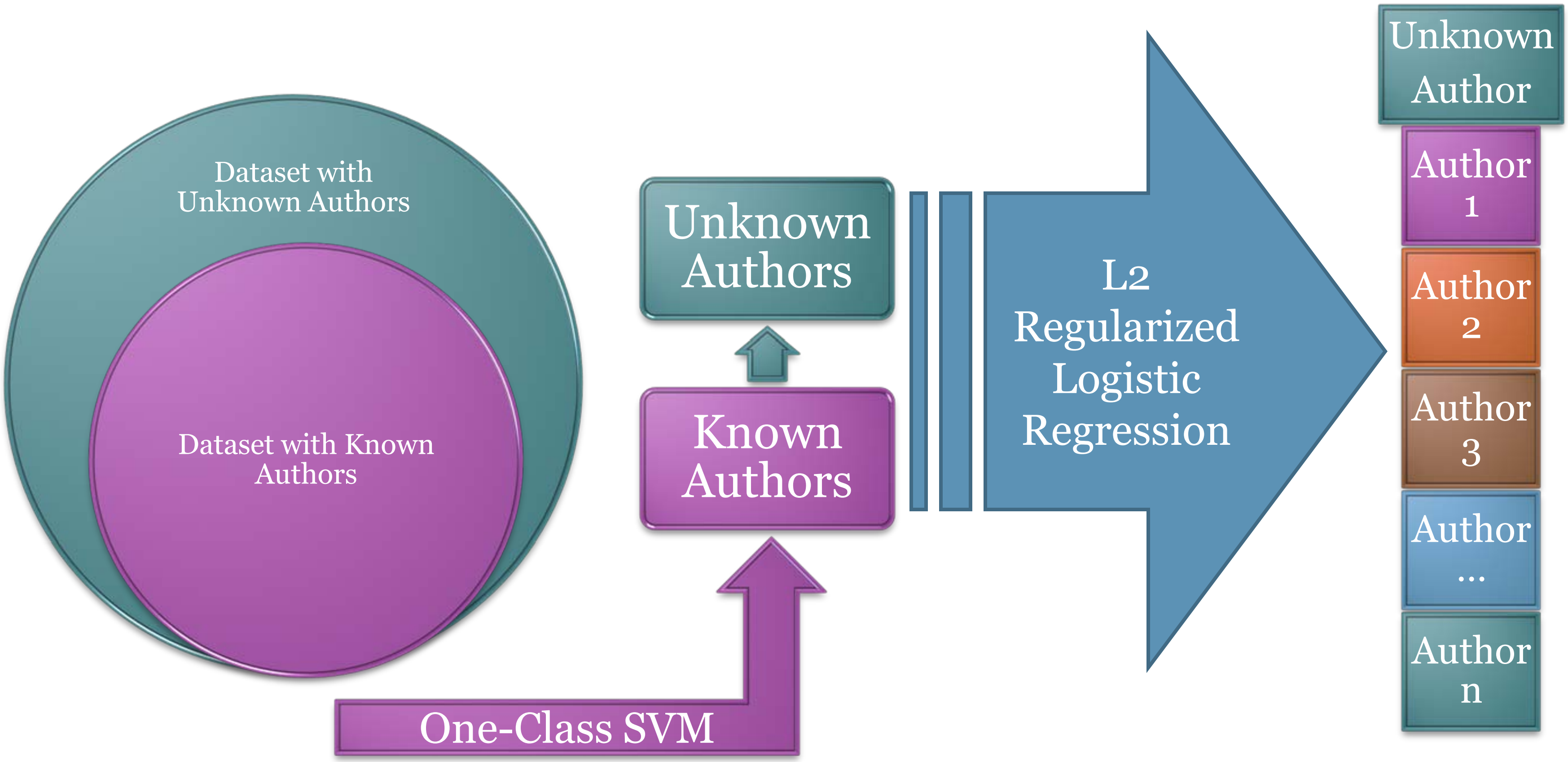
Results in Small Train Dataset



F_1 in Small Test Dataset

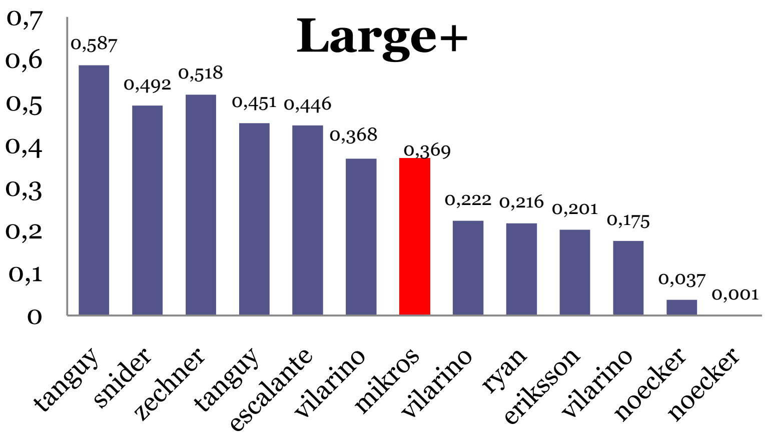


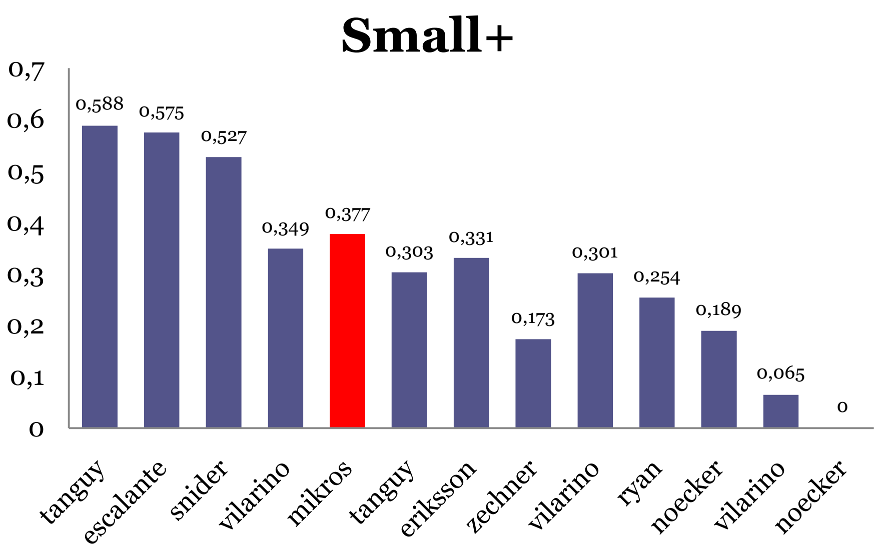
Procedure in Large & Small + Datasets



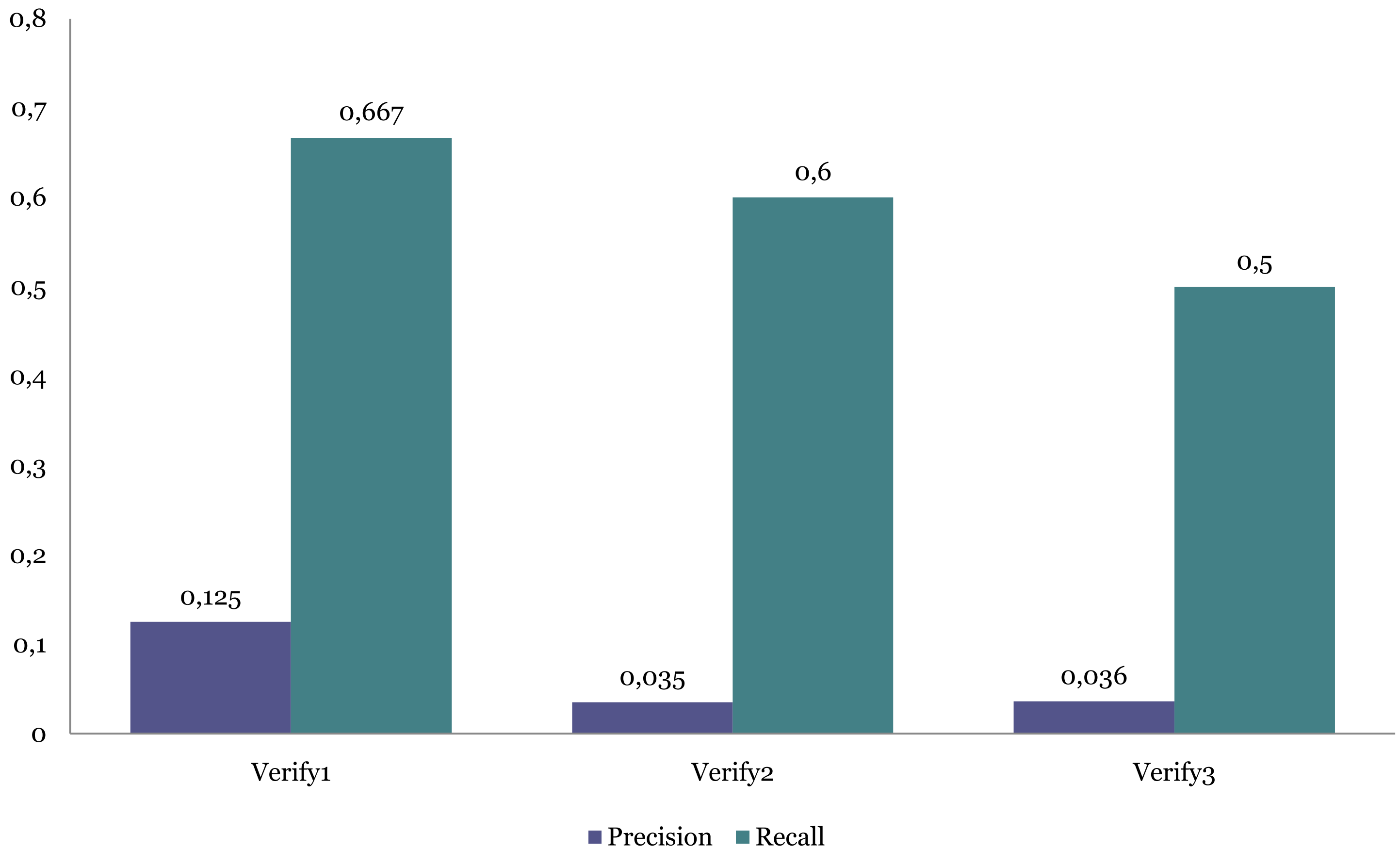
PAN 2011 Lab,
19-22 September
2011, Amsterdam

F_1 in Large & Small +





Results in Verification datasets



Conclusions

- Features spanning in multiple linguistic levels capture better author's stylistic variation than features that focus in a specific level.
- L2 Regularized Logistic Regression performs very well in high dimensional data.
- Authorship verification research remains a difficult problem and research should be focused to new algorithms handling one-class problems.
- We need one / many common benchmark corpus/corpora in order to further advance authorship identification tools and methods.