Authorship Attribution in Greek Tweets Using Author's Multilevel N-gram Profiles

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

A brief typology of authorship research

- **Authorship attribution**: Closed problem. We assume that one of 1, 2, 3… n candidates is the real author of a text.
- **Author verification**: Open problem. We assume an open set of authors and each text should be attributed to its real author without reference to any corpus from other authors.
- **Author profiling**: Closed problem. We assume that specific extralinguistic characteristics (gender, age, psychological profile etc.) of the author(s) can be traced in his/her texts.

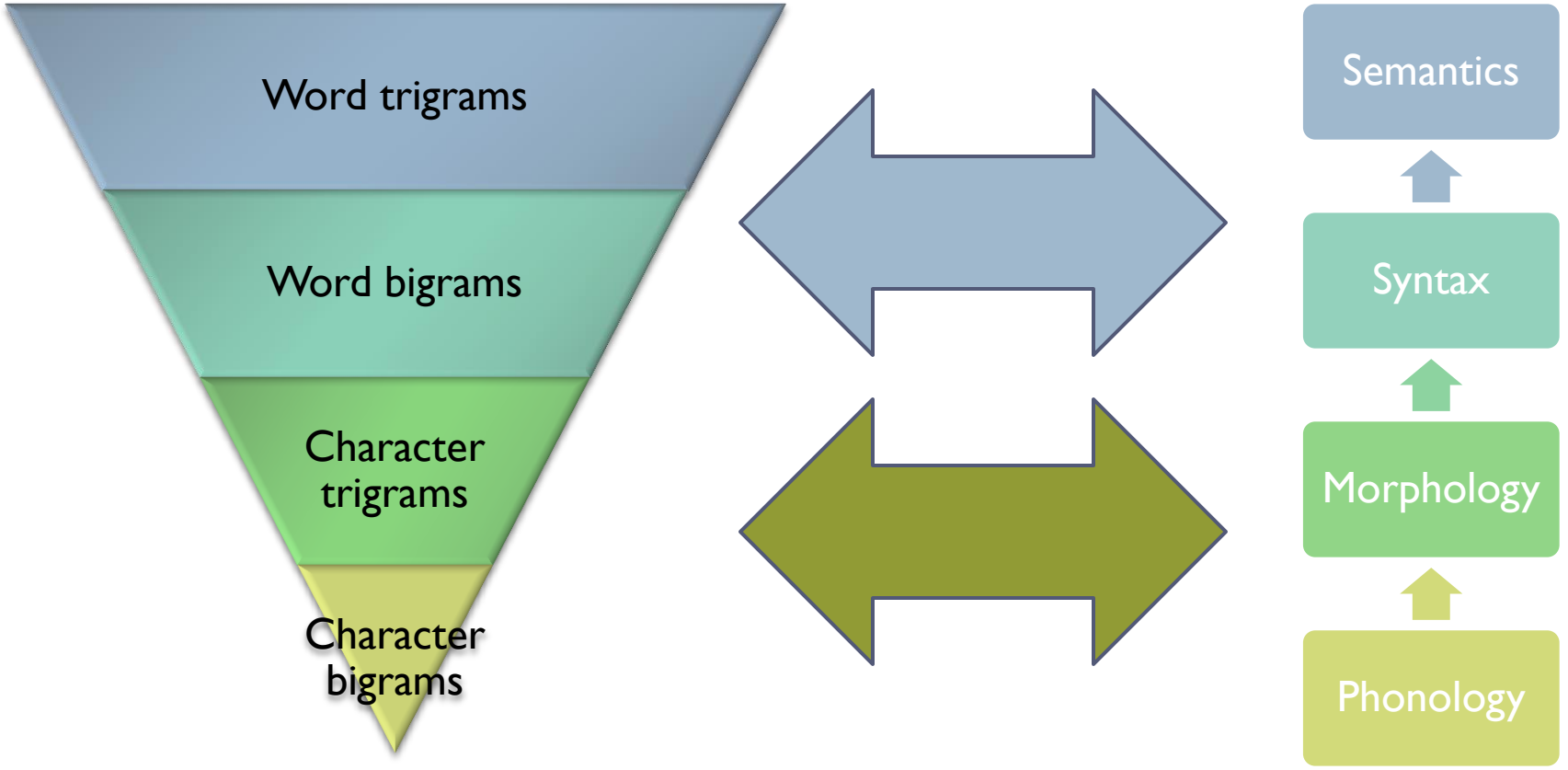
Aims of the present research

- To perform authorship attribution experiments in tweets written in Modern Greek.
  - Create the first Modern **G**reek **T**weets **C**orpus (GTC) in order to use it as a reference corpus for studying social media language including authorship attribution, sentiment analysis and linguistic variation.
  - Explore the effectiveness of a specific document representation called **Author's Multilevel N-gram Profile (AMNP)**, which comprises of a combined vector of increasing size and different level n-grams
  - Investigate alternative ways to construct training sets for authorship attribution in Twitter data. More specifically:
    - Is it better to use single tweets for testing or do we need to merge tweets producing bigger text units?
    - In the case of merging tweets, what is the text size that produces the best attribution results?

The Greek Twitter Corpus

Authors	No of Tweets	Total size (words)	Average size (words)	Standard Deviation
A	500	5,378	10.75	5.42
B	918	10,515	11.45	5.52
C	2,065	32,098	15.54	6.73
D	455	7,451	16.57	5.48
E	1,347	9,822	7.29	5.01
F	535	3,692	6.90	4.93
G	1,277	9,412	7.37	5.63
H	2,306	26,212	11.36	5.86
I	2,986	18,720	6.26	4.28
J	584	7,618	13.06	6.74
Total	**12,973**	**130,918**		

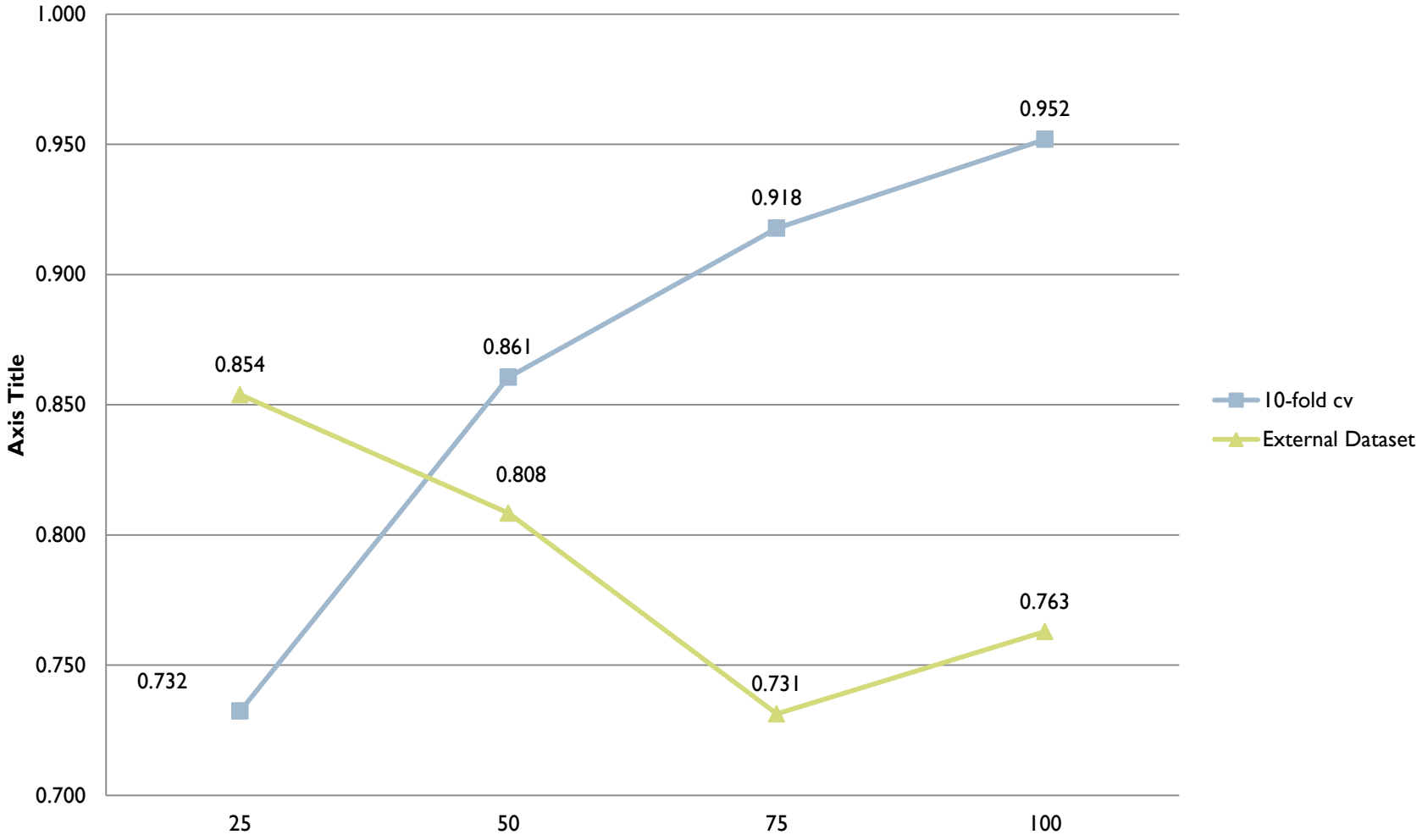
Author's Multilevel N-gram Profile - AMNP



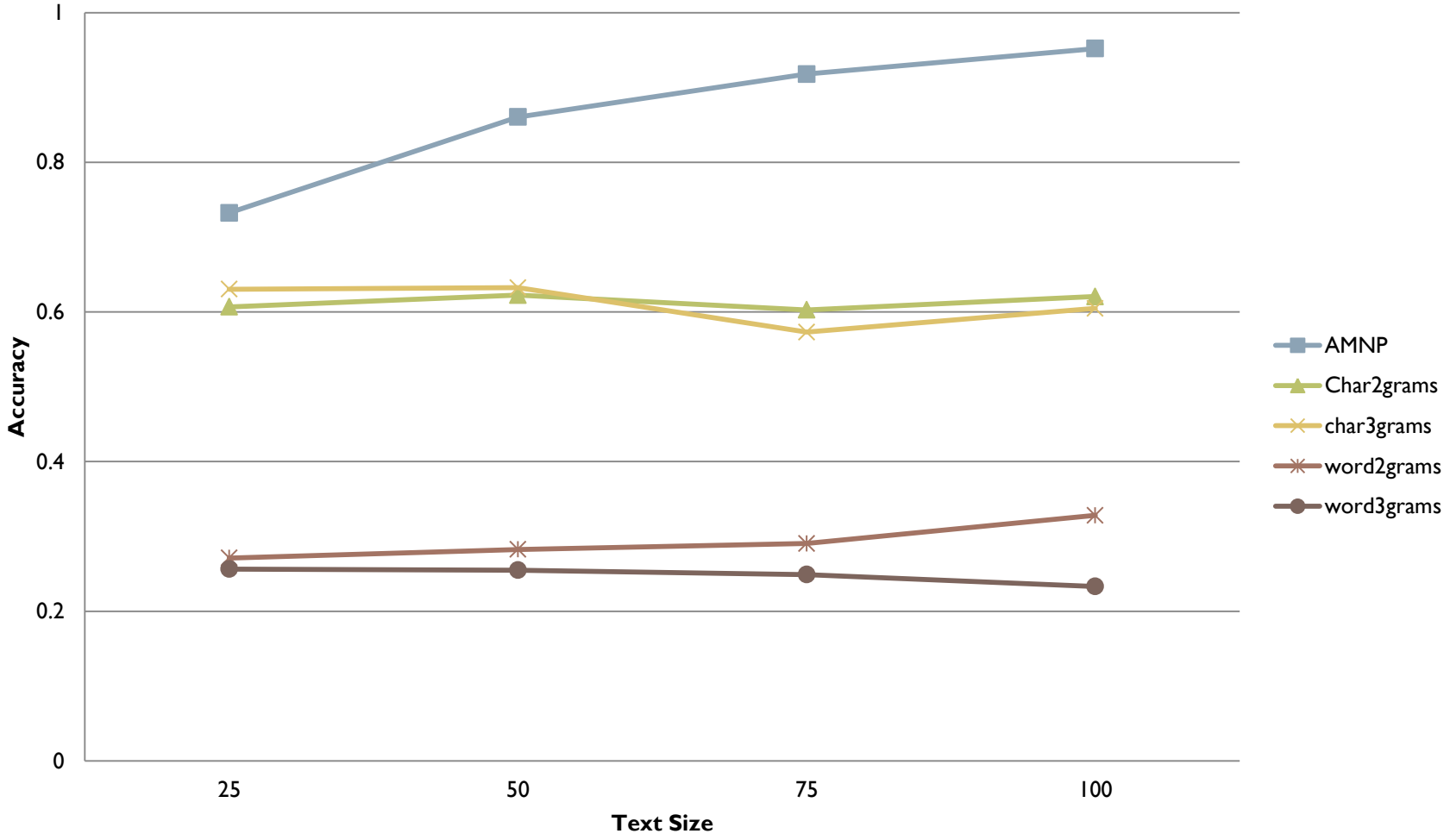
Experimental methodology

- We used GTC to create 4 different datasets which contained merged tweets of increasing text sizes (25, 50, 75 and 100 words).
- As classification algorithm we used multiclass support vector classification (LIBLINEAR library).
- We tested the authorship attribution accuracy with each feature group separately and compared it with AMNP. Accuracy figures were calculated on two different conditions:
  - a) 10-fold cross-validation (cv) in the merged tweets text units
  - b) External test set which contained 500 single tweets not included in the training set (35-60 per author).

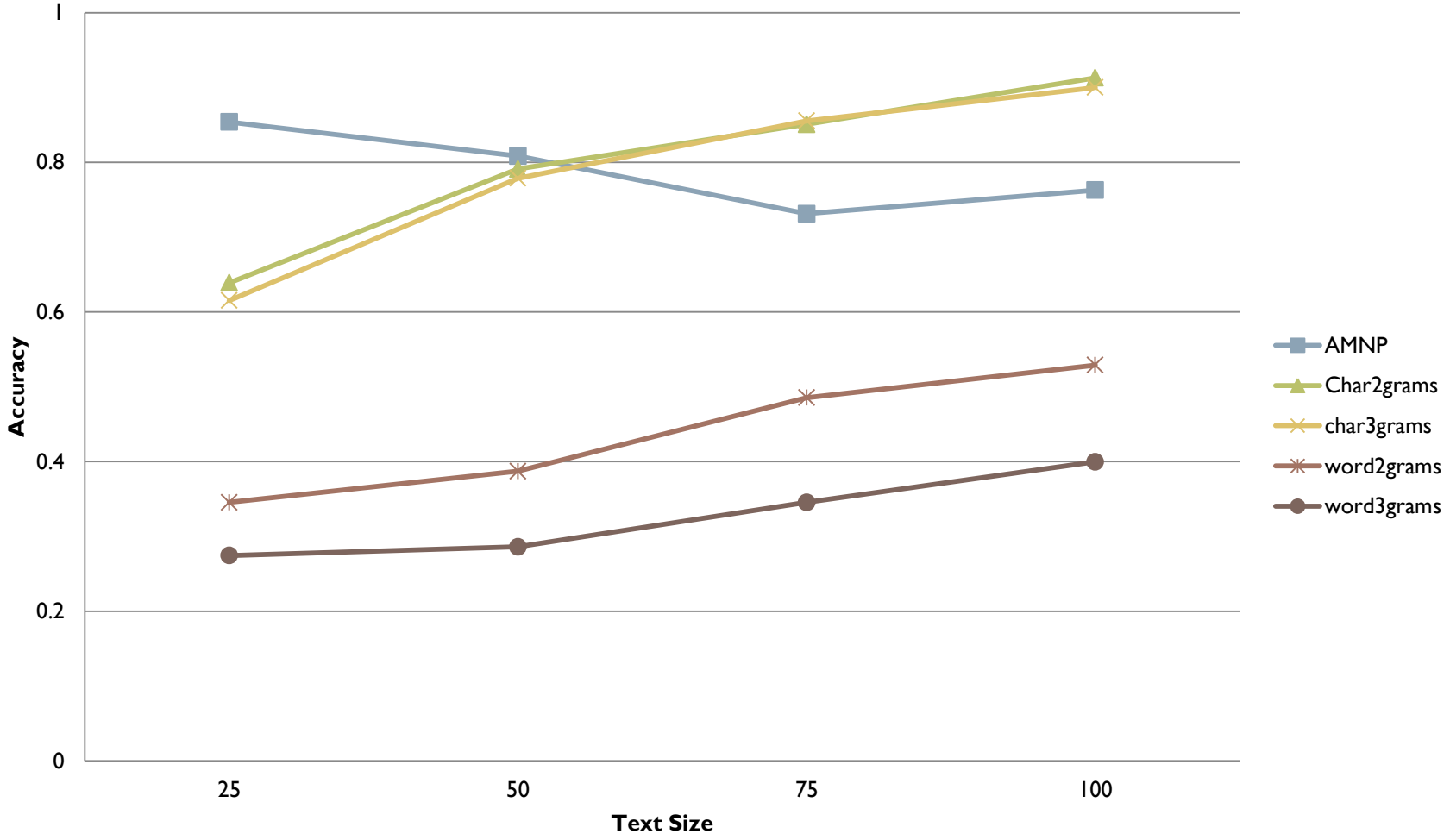
Accuracy vs. text size chunks using two different validation methods



Impact of text size and feature representation method in authorship attribution accuracy using cross-validation



Impact of text size and feature representation method in authorship attribution accuracy using external dataset



Conclusions

- Authorship attribution in tweets of Modern Greek is a feasible task. Our top performance (0.952 accuracy in 10-fold cv using 100-word text chunks) is a good indication that the tweet's linguistic structure is a significant carrier of authorship information.
- AMNP representation is based on a solid linguistic – semiotic theoretical background and proved highly efficient compared to single n-gram feature groups in all text sizes.
- The obtained results indicated that optimal performance is achieved when **both** training and testing sets for authorship attribution contained merged tweets.

Thank you…