Author vs. Translator
Using Author Multilevel Ngram Profiles for detecting both author and translator in literary texts

George K. Mikros

gmikros@isll.uoa.gr

National and Kapodistrian University of Athens, Greece

University of Massachusetts Boston, USA

Overview

- Authorship Identification (AuI) premises
- Extending AuI methods to Translator Attribution
- Research aims of the study
- Experimental Methodology
  - Corpus creation
  - Features (AMNP)
  - Machine learning classification algorithms (SVM & RF)
  - Model evaluation
  - Features’ evaluation
- Conlusions

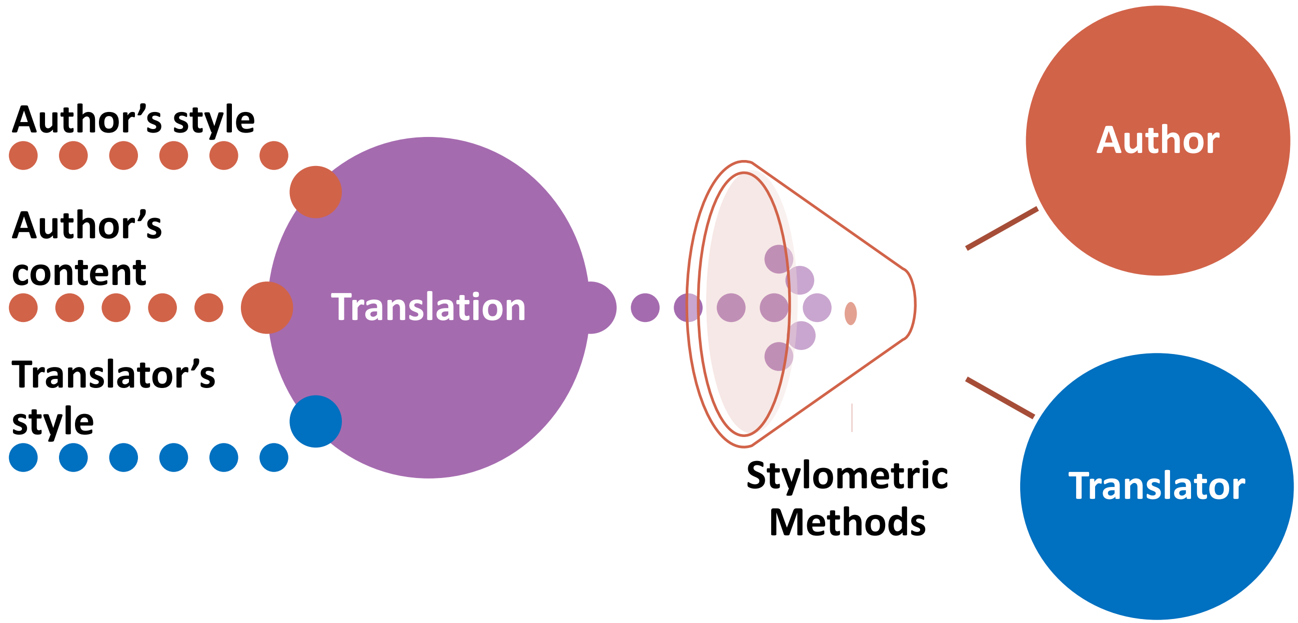
Authorship Identification (AuI) studies: a brief typology

- Authorship identification refers to the connection of a text of unknown authorship to a specific author (or author characteristics) using a set of quantifiable text features as indicators of the author's style.
  - **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.

Extending AuI methodology: Translator Attribution

- **Premises**: Stylometric theory assumes that each author possess a distinct, unique "writeprint" which is expressed quantitatively through the idiosyncratic occurrence variation of its most frequent linguistic structures and various indices of unconscious linguistic behavior such as lexical "richness" formulas, word and sentence lengths etc.
- **Translations**: The ultimate test of "writeprint" theory.
  - Translator's attribution gives evidence that:
    - Each human has a distinct stylometric "signature", which is detectable even when someone translates a text written in different language and by a different author.
    - Stylometric methods can capture deep cognitive aspects of linguistic identities that pertain across language codes, content and text genres.

Translator attribution as signal decomposition

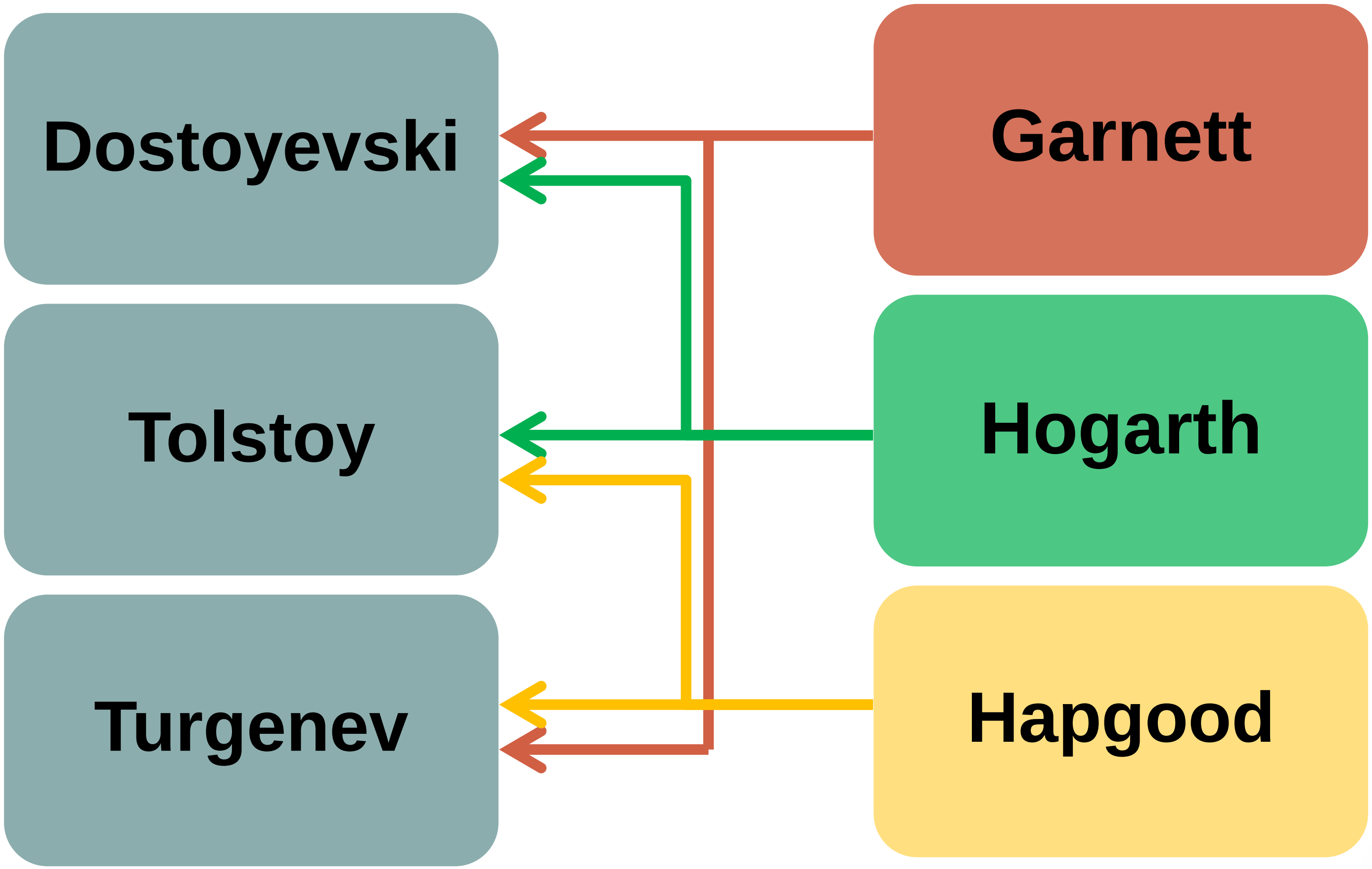


Corpus compilation

Author	Translator	Title	Words
Fyodor Dostoyevski	**Constance Garnett**	The Brothers Karamazov	359,490
Fyodor Dostoyevski	**Constance Garnett**	Crime and punishment	204,267
Fyodor Dostoyevski	**Hogarth, C. J.**	Poor Folk	54,866
Fyodor Dostoyevski	**Hogarth, C. J.**	The Gambler	61,068
Leo Tolstoy	**Hapgood, Isabel Florence**	The Census in Moscow	4,241
Leo Tolstoy	**Hogarth, C. J.**	Boyhood	28,843
Leo Tolstoy	**Hogarth, C. J.**	Childhood	39,005
Turgenev, Ivan	**Constance Garnett**	A House of Gentlefolk	62,115
Turgenev, Ivan	**Hapgood, Isabel Florence**	A Reckless Character	81,017

In order to increase our sample space and create enough data points for valid statistical measurements we segmented each text in **1,000** word chunks, creating a dataset of **879** texts.

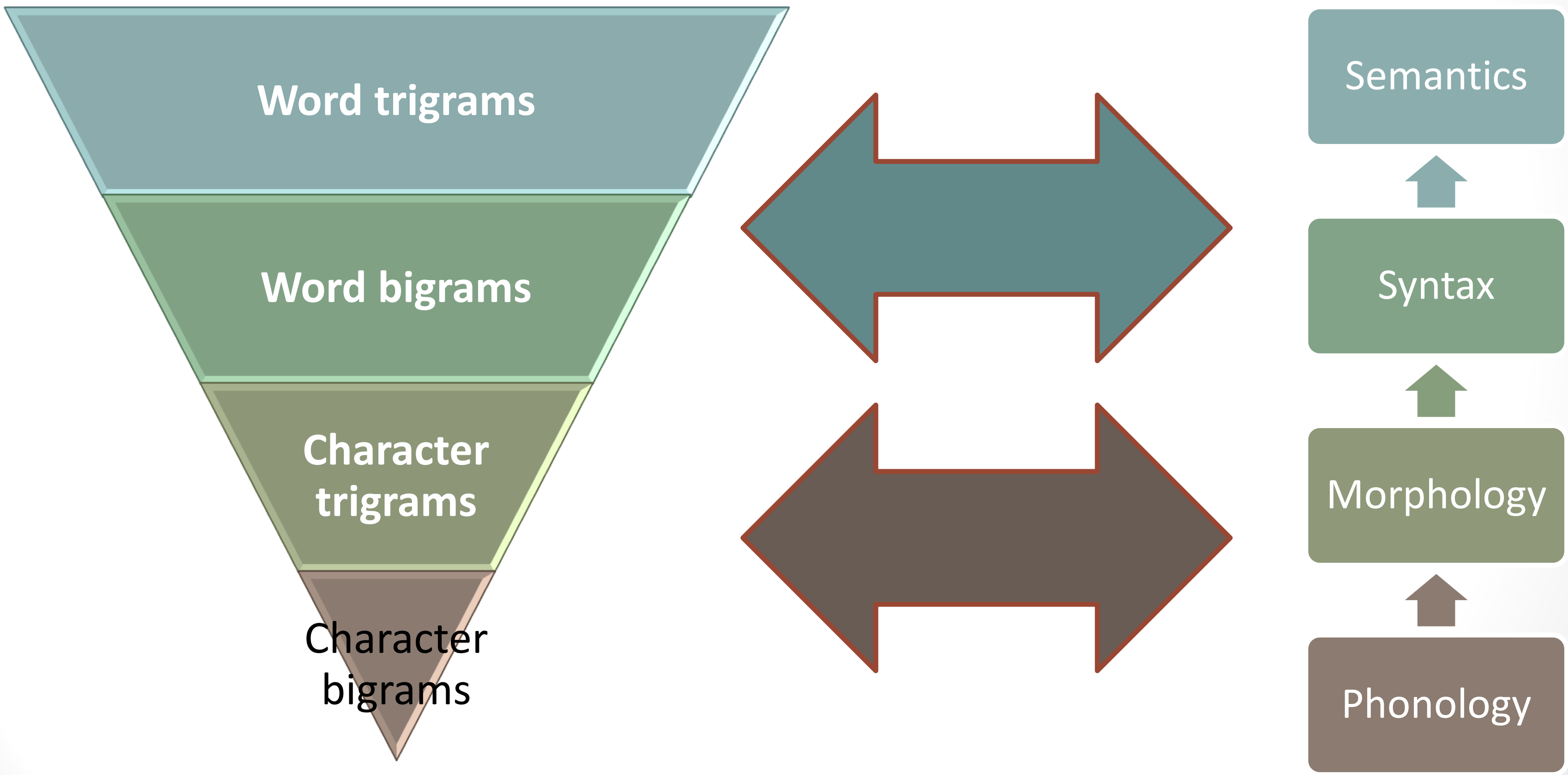
Authors ~ Translators correspondance



Feature representation: N-grams

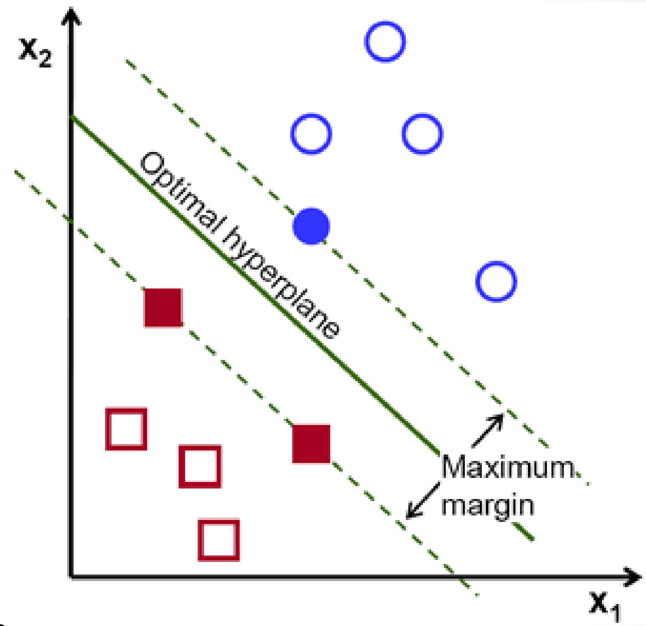
- Character and word n-grams have been used successfully previously in AAI tasks with character bigrams to appear as early as 1976 in the relative literature (Bennett 1976).
- They exhibit significant advantages over other stylometric features since their identification can be achieved easily and they are language-independent.
- Taking into consideration the complementary nature of character and word level information, we propose a combined vector of both character and word n-grams of different size.
- The resulting vector represents the **Author's Multilevel N-gram Profile (AMNP)**, a document representation that captures in a parallel way both character and word sequences.
- Using AMNP we combine information from different linguistic levels and we capture stylistic variation across a wide range of linguistic choices.

Author's Multilevel N-gram Profile - AMNP



Support Vector Machines

- A **support vector machine** (**SVM**) is a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis (Vapnik 1995).
- It involves finding the hyperplane (line in 2D, plane in 3D and hyperplane in higher dimensions.
- More formally, a hyperplane is n-1 dimensional subspace of an n-dimensional space) that best separates two classes of points with the maximum margin.
- The data points that kind of "support" this hyperplane on either sides are called the "support vectors".
- For cases where the two classes of data are not linearly separable, the points are projected to an exploded (higher dimensional) space where linear separation may be possible.



Random Forests

- A **random forest** is an ensemble (i.e., a collection) of unpruned decision trees (Breiman 2001).
- Random forests are often used when we have very large training datasets and a very large number of input variables (hundreds or even thousands of input variables). A random forest model is typically made up of tens or hundreds of decision trees.
- Can be used for classification or regression.
- Accuracy and variable importance information is provided with the results.

Experimental procedure

- Two experiments:
  - Authorship attribution:
    - 3 authors: Dostoyevski, Tolstoy, Turgenev
  - Translator attribution:
    - 3 translators: Garnett, Hogarth, Hapgood
- Corpus: All translations. Splitting in training set (75% of the original corpus) and testing set (25%).
- Features: AMNP
  - 2,000 (500 most frequent n-grams from each n-gram category (character 2-grams, 3-grams, word 2-grams, 3-grams).
  - Feature reduction due to data sparsity using near-zero variance predictor detection (1,607 n-grams).
    - The percentage of unique values is less than 20% and
    - The ratio of the most frequent to the second most frequent value is greater than 20
- Classification algorithm: SVM (polynomial kernel) and RF.
- Parameter tuning: 3 points grid-search
- Models Training: Parameter estimation using 10-fold cross-validation in the training set
- Models Evaluation: Accuracy on the testing set (25% of the original corpus).

Authorship Attribution results

SVM model training data: 0.98 10-fold cv accuracy

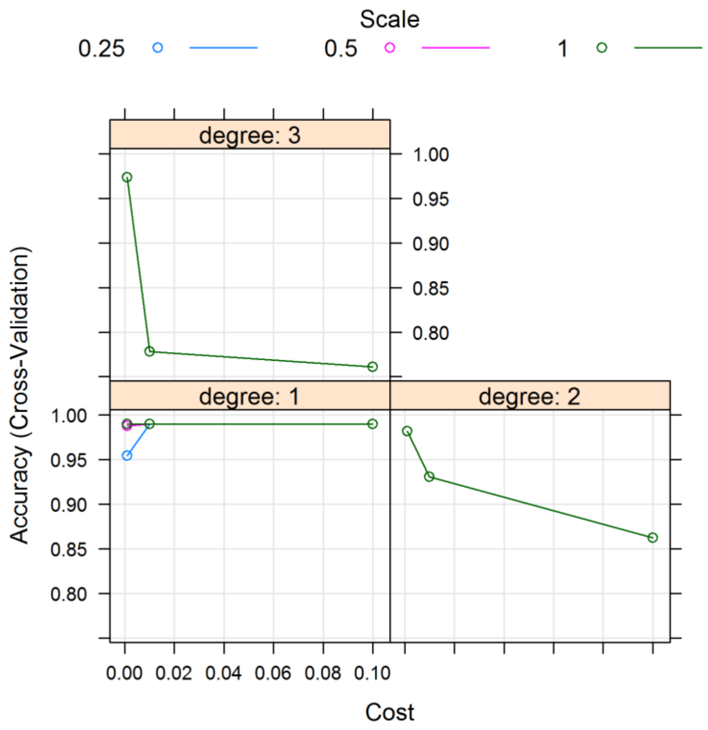
	Reference		
Prediction	Dostoyefski	Tolstoy	Turgenev
Dostoyefski	**500**	6	1
Tolstoy	0	**47**	1
Turgenev	1	0	**104**

RF model training data: 0.90 10-fold cv accuracy

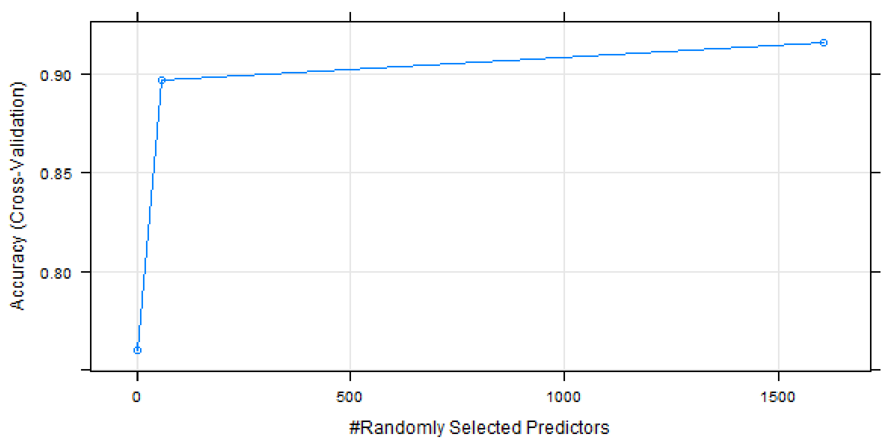
	Reference		
Prediction	Dostoyefski	Tolstoy	Turgenev
Dostoyefski	**498**	30	29
Tolstoy	1	**23**	1
Turgenev	2	0	**76**

Authorship Attribution tuning process

SVM tuning process



RF tuning process



Chicago Colloquium on Digital Humanities and Computer Science
DHCS 14, October 23-24, 2014

Translator Attribution results

SVM model training data: 0.99 10-fold cv accuracy

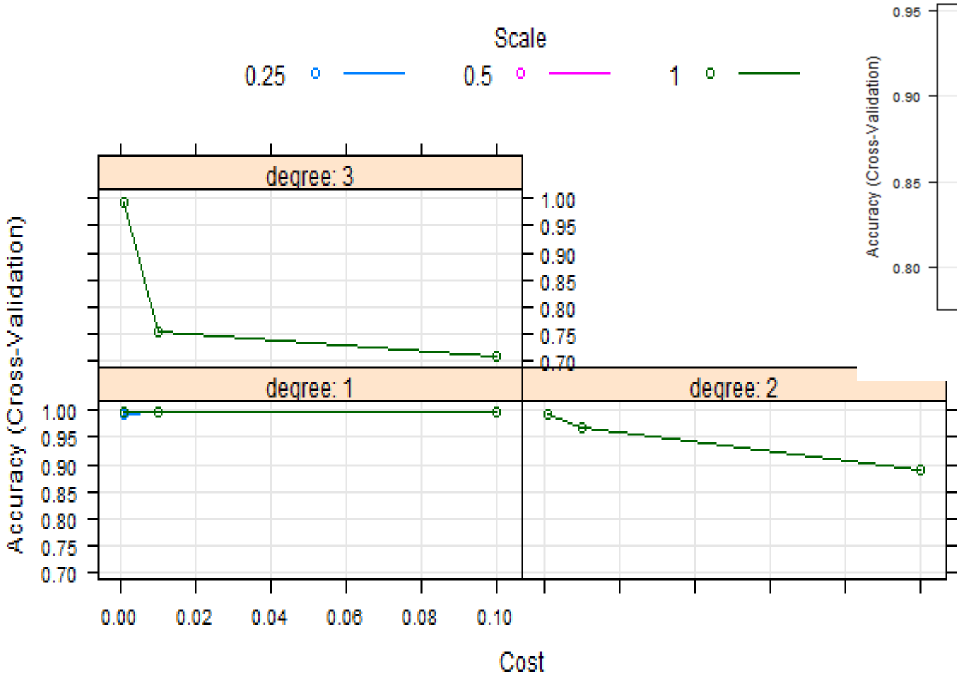
	Reference		
Prediction	Garnett	Hapgood	Hogarth
Garnett	**461**	1	0
Hapgood	0	**62**	0
Hogarth	1	0	**135**

RF model training data: 0.92 10-fold cv accuracy

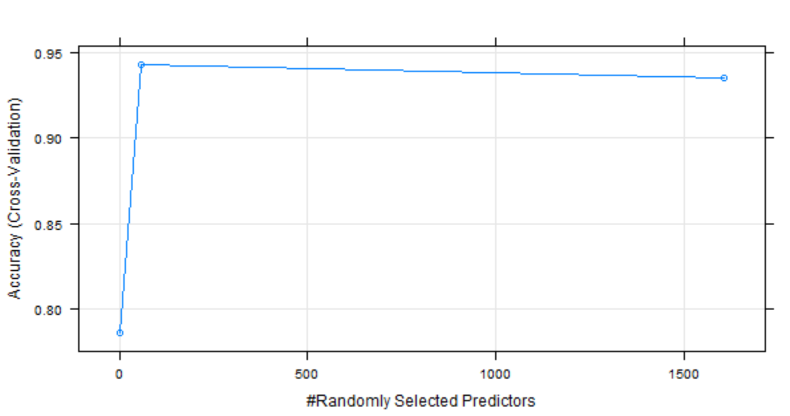
	Reference		
Prediction	Garnett	Hapgood	Hogarth
Garnett	**460**	35	13
Hapgood	0	**28**	1
Hogarth	2	1	**121**

Translator Attribution tuning process

SVM tuning process



RF tuning process



Model evaluation in the testing data

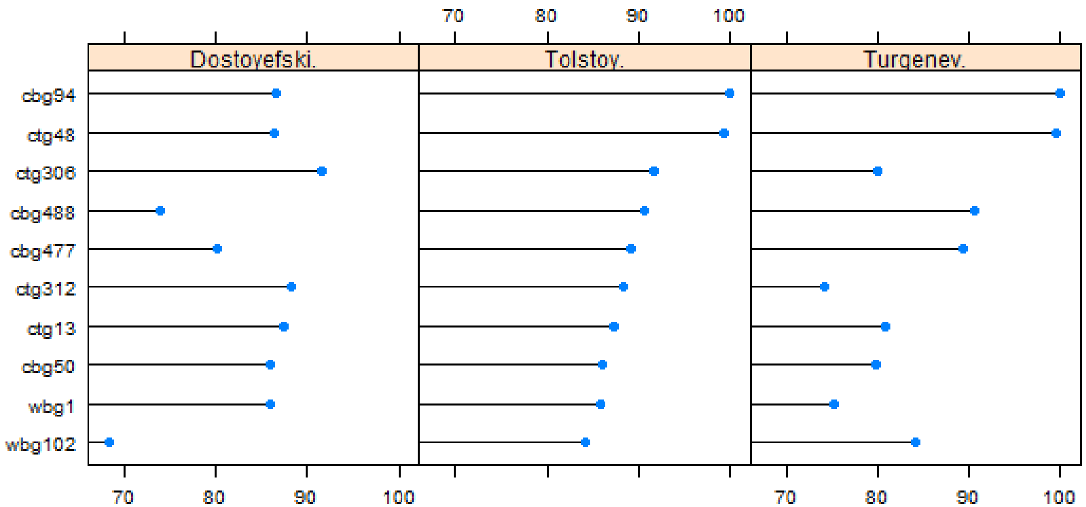
Authorship SVM model testing data: 0.99 accuracy

	Reference		
Prediction	Dostoyefski	Tolstoy	Turgenev
Dostoyefski	**167**	0	0
Tolstoy	2	**15**	0
Turgenev	0	0	**35**

Translator SVM model testing data: 1 accuracy

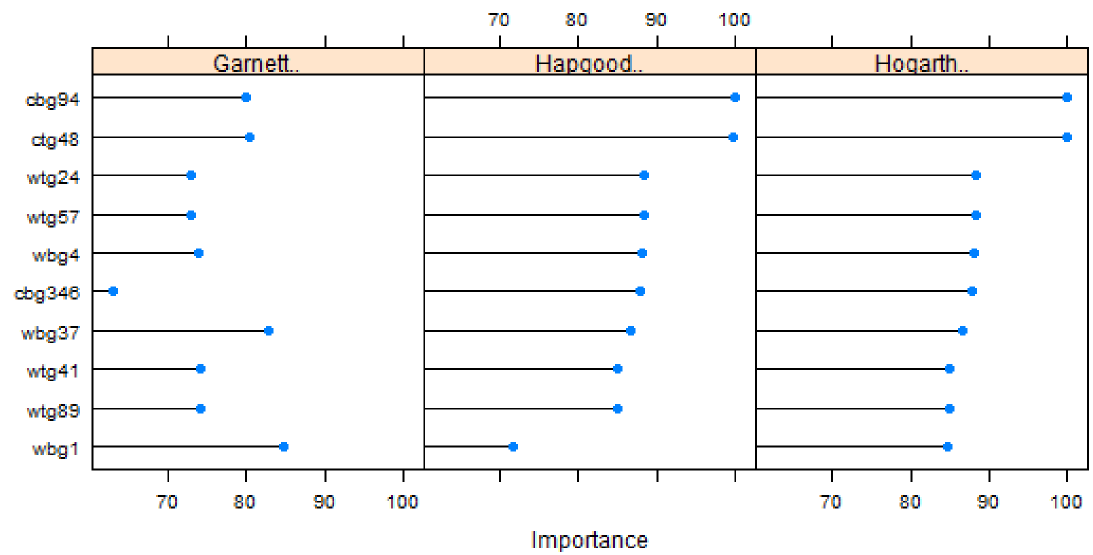
	Reference		
Prediction	Garnett	Hapgood	Hogarth
Garnett	**167**	0	0
Hapgood	0	**17**	0
Hogarth	0	0	**35**

Features' importance (area under ROC)

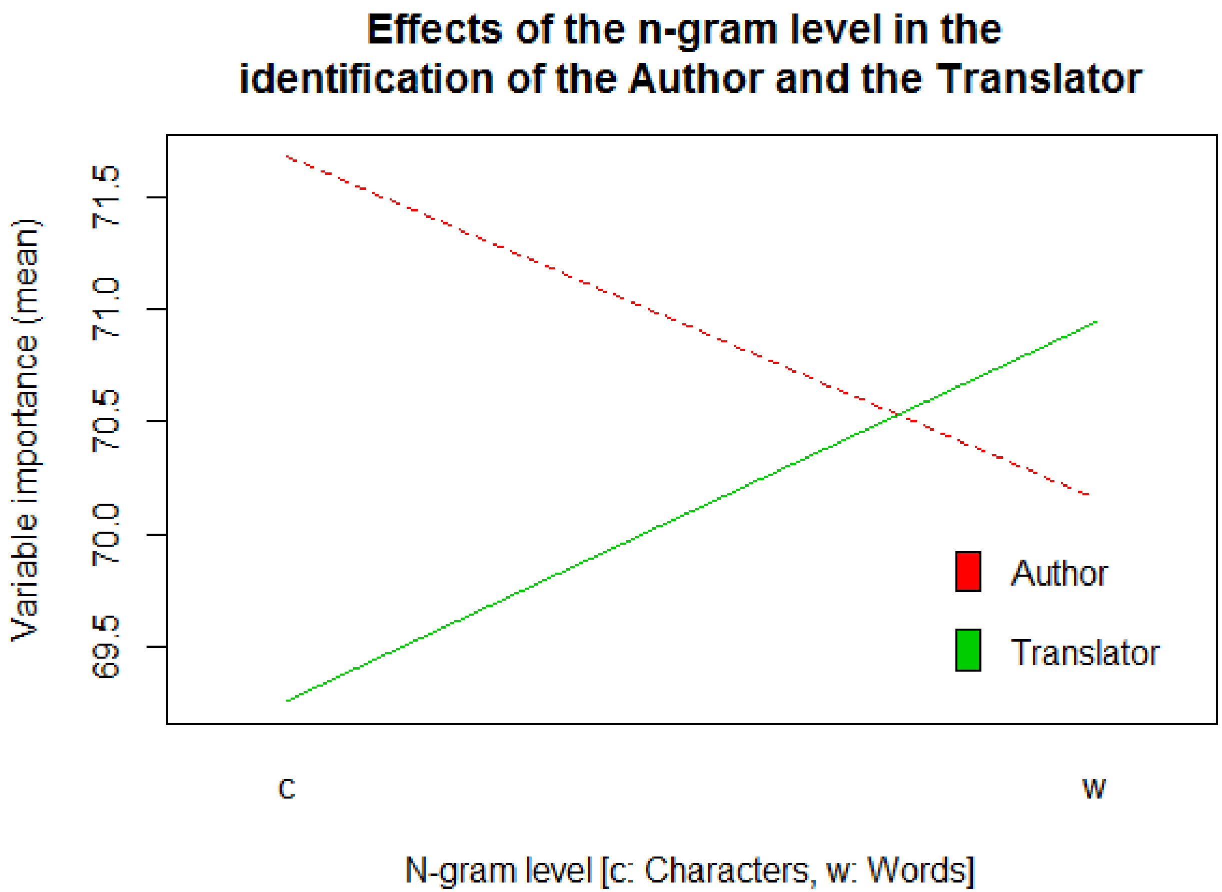


Authorship Attribution

Translator Attribution



N-gram level effect on Author and Translator identification



Two-way ANOVA statistically significant ($p<0.001$)

Conclusions

The reported experiments tried to explore the possibility to apply authorship attribution techniques to the translator identification problem. Our results suggest that:

- **Author and Translator attribution is feasible** with high accuracy in small closed class groups of candidate authors and translators.
- **AMNP seems to be a promising document representation** methodology especially in problems where the attribution requires uncovering subtle differences in linguistic usage.
- **SVMs are combined better with AMNP** in these dual aim classifications (author ~ translator) due to their ability to create higher-order hyperplanes embedding subsets of n-grams depending on the classification aim.
- **Translator is not “invisible”.**
  - Word n-grams seem to convey stylistic choices of the translator.
  - Character n-grams provide authorial information.
- Future work should be directed to controlled experiments of author vs. translator problems with more candidates and research on cases where the author is at the same time translator.

Thank you!