From Language to Family and Back
Native Language and Language Family Identification from English Text

Ariel Stolerman    Aylin Caliskan Islam    Rachel Greenstadt

Dept. of Computer Science
Drexel University
Philadelphia, PA

NAACL HLT SRW 2013
Outline

1. Motivation
   - Problem Statement
   - Related Work

2. Contribution
   - Experimental Settings
   - Experimental Variations

3. Conclusion
Outline

1 Motivation
   - Problem Statement
   - Related Work

2 Contribution
   - Experimental Settings
   - Experimental Variations

3 Conclusion
Outline

1. Motivation
   - Problem Statement
   - Related Work

2. Contribution
   - Experimental Settings
   - Experimental Variations

3. Conclusion
Outline

1 Motivation
   - Problem Statement
   - Related Work

2 Contribution
   - Experimental Settings
   - Experimental Variations

3 Conclusion
Stylometry: the study of linguistic style
Question: Who wrote this document?
Utilize stylometric analysis to characterize anonymous authors
- Age, Gender, Native language...
- Profiling criminals / terrorists / unveil identities
Problem Statement and Motivation

- Stylometry: the study of linguistic style
- Question: *Who wrote this document?*
- Utilize stylometric analysis to characterize anonymous authors
  - Age, Gender, *Native language*...
  - Profiling criminals / terrorists / unveil identities
Problem Statement and Motivation

▶ Stylometry: the study of linguistic style
Question: *Who wrote this document?*
▶ Utilize stylometric analysis to characterize anonymous authors
  ▶ Age, Gender, *Native language*...
  ▶ Profiling criminals / terrorists / unveil identities
Stylometry: the study of linguistic style
Question: *Who wrote this document?*

Utilize stylometric analysis to characterize anonymous authors
- Age, Gender, *Native language*...
- Profiling criminals / terrorists / unveil identities
Definitions

- **L1**: native language
- **L2**: non-native language
- **LF**: language family of L1

**Question of interest:**
Given English text by an unknown author, what is his/her native language(s)?

- **L1-L2 transfer effect** $\rightarrow$ **LF-L2 transfer effect**?
- **Increase L1-ID via LF-ID**?
  - Yes
Problem Statement and Motivation

Definitions
- L1: native language
- L2: non-native language
- LF: language family of L1

Question of interest:
*Given English text by an unknown author, what is his/her native language(s)?*
- L1-L2 transfer effect → LF-L2 transfer effect?
- Increase L1-ID via LF-ID?
  - Yes
Problem Statement and Motivation

- Definitions
  - L1: native language
  - L2: non-native language
  - LF: language family of L1

- Question of interest:
  Given English text by an unknown author, what is his/her native language(s)?
  - L1-L2 transfer effect → LF-L2 transfer effect?
  - Increase L1-ID via LF-ID?
    - Yes

Definitions

- L1: native language
- L2: non-native language
- LF: language family of L1

Question of interest:
Given English text by an unknown author, what is his/her native language(s)?
- L1-L2 transfer effect → LF-L2 transfer effect?
- Increase L1-ID via LF-ID?
  - Yes
Problem Statement and Motivation

Definitions

- L1: native language
- L2: non-native language
- LF: language family of L1

Question of interest:
Given English text by an unknown author, what is his/her native language(s)?

- L1-L2 transfer effect → LF-L2 transfer effect?
- Increase L1-ID via LF-ID?
  - Yes
Outline

1. Motivation
   - Problem Statement
   - Related Work

2. Contribution
   - Experimental Settings
   - Experimental Variations

3. Conclusion
Related Work

- Choueiter et al.: L1 and accent identification from speech [CZN08]
- L1 identification from text
  - Koppel et al. [KSZ05a, KSZ05b]
  - [EGP+07, vH08, WDJ11, BH12]...
- None consider LF effects and utilization for L1-ID
Related Work

▶ Choueiter et al.: L1 and accent identification from speech [CZN08]
▶ L1 identification from text
  ▶ Koppel et al. [KSZ05a, KSZ05b]
  ▶ [EGP+07, vH08, WDJ11, BH12]...
▶ None consider LF effects and utilization for L1-ID
Outline

1. Motivation
   - Problem Statement
   - Related Work

2. Contribution
   - Experimental Settings
   - Experimental Variations

3. Conclusion
Corpus, Features, Classifier

- **Corpus:** ICLEv2
  - English texts by international learners of English, 16 mother-tongues
  - Used 11 L1 of 3 LF: Romance, Slavic & Germanic

- **Feature Sets**
  - **Basic:** Adopted from Koppel et al. [KSZ05a, KSZ05b]
    - Function words, frequent letter bigrams, rare part-of-speech bigrams, common misspellings
  - **Extended:** added grammar features
    - Frequent POS bigrams
  - **Grammar:** only grammar features
  - **InfoGain:** Feature space reduction on Extended

- Measured TPR with SVM classifier ± cross-validation
Corpus, Features, Classifier

- **Corpus: ICLEv2**
  - English texts by international learners of English, 16 mother-tongues
  - Used 11 L1 of 3 LF: Romance, Slavic & Germanic

- **Feature Sets**
  - **Basic**: Adopted from Koppel et al. [KSZ05a, KSZ05b]
    - Function words, frequent letter bigrams, rare part-of-speech bigrams, common misspellings
  - **Extended**: added grammar features
    - Frequent POS bigrams
  - **Grammar**: only grammar features
  - **InfoGain**: Feature space reduction on Extended
  - Measured TPR with SVM classifier + cross-validation
Corpus, Features, Classifier

- **Corpus:** ICLEv2
  - English texts by international learners of English, 16 mother-tongues
  - Used 11 L1 of 3 LF: Romance, Slavic & Germanic

- **Feature Sets**
  - **Basic:** Adopted from Koppel et al. [KSZ05a, KSZ05b]
    - Function words, frequent letter bigrams, rare part-of-speech bigrams, common misspellings
  - **Extended:** added grammar features
    - Frequent POS bigrams
  - **Grammar:** only grammar features
  - **InfoGain:** Feature space reduction on Extended

- Measured TPR with SVM classifier + cross-validation
Corpus, Features, Classifier

- **Corpus**: ICLEv2
  - English texts by international learners of English, 16 mother-tongues
  - Used 11 L1 of 3 LF: Romance, Slavic & Germanic

- **Feature Sets**
  - **Basic**: Adopted from Koppel et al. [KSZ05a, KSZ05b]
    - Function words, frequent letter bigrams, rare part-of-speech bigrams, common misspellings
  - **Extended**: added grammar features
    - Frequent POS bigrams
  - **Grammar**: only grammar features
  - **InfoGain**: Feature space reduction on Extended

- Measured TPR with SVM classifier + cross-validation
Corpus, Features, Classifier

- Corpus: ICLEv2
  - English texts by international learners of English, 16 mother-tongues
  - Used 11 L1 of 3 LF: Romance, Slavic & Germanic

- Feature Sets
  - **Basic**: Adopted from Koppel et al. [KSZ05a, KSZ05b]
    - Function words, frequent letter bigrams, rare part-of-speech bigrams, common misspellings
  - **Extended**: added grammar features
    - Frequent POS bigrams
  - **Grammar**: only grammar features
  - **InfoGain**: Feature space reduction on **Extended**

- Measured TPR with SVM classifier + cross-validation
Corpus, Features, Classifier

- **Corpus: ICLEv2**
  - English texts by international learners of English, 16 mother-tongues
  - Used 11 L1 of 3 LF: Romance, Slavic & Germanic

- **Feature Sets**
  - **Basic**: Adopted from Koppel et al. [KSZ05a, KSZ05b]
    - Function words, frequent letter bigrams, rare part-of-speech bigrams, common misspellings
  - **Extended**: added grammar features
    - Frequent POS bigrams
  - **Grammar**: only grammar features
  - **InfoGain**: Feature space reduction on Extended

- Measured TPR with SVM classifier + cross-validation
Outline

1. Motivation
   - Problem Statement
   - Related Work

2. Contribution
   - Experimental Settings
   - Experimental Variations

3. Conclusion
9-Class L1 vs. 3-Class LF

- Questions
  - Best LF-ID method?
  - L1-ID vs. LF-ID performance on same data?
- LF-ID methods
  - Trivial
    - Apply L1-ID
    - Take LF of attributed L1
  - Standalone
    - Apply LF-ID directly (using LF as classes)
  - Combined
    - Set confidence threshold $t$
    - Apply L1-ID, measure chosen L1 probability $p$
    - If $p \geq t$: take LF by Trivial
    - If $p < t$: take LF by Standalone
    - Hypothesis: Good L1-ID $\rightarrow$ Good LF-ID from that L1
9-Class L1 vs. 3-Class LF

- Questions
  - Best LF-ID method?
  - L1-ID vs. LF-ID performance on same data?
- LF-ID methods
  - Trivial
    - Apply L1-ID
    - Take LF of attributed L1
  - Standalone
    - Apply LF-ID directly (using LF as classes)
  - Combined
    - Set confidence threshold $t$
    - Apply L1-ID, measure chosen L1 probability $p$
    - If $p \geq t$: take LF by Trivial
    - If $p < t$: take LF by Standalone
    - Hypothesis: Good L1-ID $\rightarrow$ Good LF-ID from that L1
9-Class L1 vs. 3-Class LF

- **Questions**
  - Best LF-ID method?
  - L1-ID vs. LF-ID performance on same data?
- **LF-ID methods**
  - **Trivial**
    - Apply L1-ID
    - Take LF of attributed L1
  - **Standalone**
    - Apply LF-ID directly (using LF as classes)
  - **Combined**
    - Set confidence threshold $t$
    - Apply L1-ID, measure chosen L1 probability $p$
    - If $p \geq t$: take LF by Trivial
    - If $p < t$: take LF by Standalone
    - Hypothesis: Good L1-ID $\rightarrow$ Good LF-ID from that L1
9-Class L1 vs. 3-Class LF

- Questions
  - Best LF-ID method?
  - L1-ID vs. LF-ID performance on same data?
- LF-ID methods
  - Trivial
    - Apply L1-ID
    - Take LF of attributed L1
  - Standalone
    - Apply LF-ID directly (using LF as classes)
  - Combined
    - Set confidence threshold $t$
    - Apply L1-ID, measure chosen L1 probability $p$
    - If $p \geq t$: take LF by Trivial
    - If $p < t$: take LF by Standalone
    - Hypothesis: Good L1-ID $\rightarrow$ Good LF-ID from that L1
9-Class L1 vs. 3-Class LF

- Questions
  - Best LF-ID method?
  - L1-ID vs. LF-ID performance on same data?
- LF-ID methods
  - Trivial
    - Apply L1-ID
    - Take LF of attributed L1
  - Standalone
    - Apply LF-ID directly (using LF as classes)
  - Combined
    - Set confidence threshold $t$
    - Apply L1-ID, measure chosen L1 probability $p$
    - If $p \geq t$: take LF by Trivial
    - If $p < t$: take LF by Standalone
- Hypothesis: Good L1-ID $\rightarrow$ Good LF-ID from that L1
9-Class L1 vs. 3-Class LF

- Questions
  - Best LF-ID method?
  - L1-ID vs. LF-ID performance on same data?

- LF-ID methods
  - Trivial
    - Apply L1-ID
    - Take LF of attributed L1
  - Standalone
    - Apply LF-ID directly (using LF as classes)
  - Combined
    - Set confidence threshold $t$
    - Apply L1-ID, measure chosen L1 probability $p$
    - If $p \geq t$: take LF by Trivial
    - If $p < t$: take LF by Standalone
    - Hypothesis: Good L1-ID $\rightarrow$ Good LF-ID from that L1
9-Class L1 vs. 3-Class LF – Results
9-Class L1 vs. 3-Class LF – Results

Combined outperforms Trivial, Standalone and L1

- Basic
- Extended
- Grammar
- InfoGain

L1 (9)
LF trivial (3)
LF standalone (3)
LF combined (3)
L1 Random
LF Random
9-Class L1 vs. 3-Class LF – Effective Results

Effective results: $\text{diff}(\text{results}, \text{random})$

![Bar chart](chart.png)

- L1 (9)
- LF trivial (3)
- LF standalone (3)
- LF combined (3)
9-Class L1 vs. 3-Class LF – Effective Results

LF-Combined and L1 perform similarly

- Basic
  - L1 (9)
  - LF trivial (3)
  - LF standalone (3)
  - LF combined (3)
- Extended
- Grammar
- InfoGain

Stolerman, Caliskan Islam & Greenstadt
From Language to Family and Back
Questions

- Improve L1-ID with LF-ID?
- Correct L1-ID with LF information
  - Set confidence threshold $t$
  - Apply L1-ID, measure chosen L1 probability $p$
    - If $p \geq t$: take chosen L1
    - If $p < t$:
      - Apply LF-ID by Standalone / Trivial / Random
      - Reapply L1-ID only among languages in chosen LF
      - Take chosen L1
Questions

- Improve L1-ID with LF-ID?
- Correct L1-ID with LF information
  - Set confidence threshold $t$
  - Apply L1-ID, measure chosen L1 probability $p$
  - If $p \geq t$: take chosen L1
  - If $p < t$:
    - Apply LF-ID by Standalone / Trivial / Random
    - Reapply L1-ID only among languages in chosen LF
    - Take chosen L1
Questions

- Improve L1-ID with LF-ID?
- Correct L1-ID with LF information
  - Set confidence threshold $t$
  - Apply L1-ID, measure chosen L1 probability $p$
  - If $p \geq t$: take chosen L1
  - If $p < t$:
    - Apply LF-ID by Standalone / Trivial / Random
    - Reapply L1-ID only among languages in chosen LF
    - Take chosen L1
9-Class L1, Reclassify by LF

Questions
- Improve L1-ID with LF-ID?
- Correct L1-ID with LF information
  - Set confidence threshold $t$
  - Apply L1-ID, measure chosen L1 probability $p$
  - If $p \geq t$: take chosen L1
  - If $p < t$:
    - Apply LF-ID by Standalone / Trivial / Random
    - Reapply L1-ID only among languages in chosen LF
    - Take chosen L1
3.67%-6.43% increase in TPR using **Standalone** correction
Conclusion

- L1-ID can be improved by LF generalization step
  - 3.67%-6.43% increase in TPR for 9-Class L1 ($p < 0.01$)
- Adding grammar features is good for L1-ID and LF-ID
  - According to InfoGain, especially for LF-ID
  - Lexical features better for L1-ID than LF-ID
- Confident in L1 $\rightarrow$ Confident in LF of L1
  - Otherwise, find LF in a standalone experiment
Conclusion

- L1-ID can be improved by LF generalization step
  - 3.67%-6.43% increase in TPR for 9-Class L1 ($p < 0.01$)
- Adding grammar features is good for L1-ID and LF-ID
  - According to InfoGain, especially for LF-ID
  - Lexical features better for L1-ID than LF-ID
- Confident in L1 $\rightarrow$ Confident in LF of L1
  - Otherwise, find LF in a standalone experiment
L1-ID can be improved by LF generalization step
  - 3.67%-6.43% increase in TPR for 9-Class L1 ($p < 0.01$)
- Adding grammar features is good for L1-ID and LF-ID
  - According to InfoGain, especially for LF-ID
  - Lexical features better for L1-ID than LF-ID
- Confident in L1 $\rightarrow$ Confident in LF of L1
  - Otherwise, find LF in a standalone experiment
Thank You

Thank You!
Questions?

- ams573@cs.drexel.edu
- http://psal.cs.drexel.edu/
  - JStylo: authorship attribution framework
  - Anonymouth: authorship anonymization framework
Julian Brooke and Graeme Hirst.
Measuring interlanguage: Native language identification with I1-influence metrics.

Ghinwa F. Choueiter, Geoffrey Zweig, and Patrick Nguyen.
An empirical study of automatic accent classification.

Dominique Estival, Tanja Gaustad, Son B. Pham, Will Radford, and Ben Hutchinson.
Author profiling for english emails.

Automatically determining an anonymous author’s native language.

Determining an author’s native language by mining a text for errors.
In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, KDD ’05*, pages 624–628, New York, NY, USA, 2005. ACM.
For Further Reading II

Hans van Halteren. 
Source language markers in europarl translations. 

Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson.
Topic modeling for native language identification. 
Questions
- L1-ID vs. LF-ID performance in same settings?
- Real LF better than Random LF?

Compared
- 3-class L1-ID
- 3-class LF-ID (Standalone), real LF
- 3-class LF-ID (Standalone), randomly constructed LF
3-Class L1 vs. 3-Class LF

Questions
- L1-ID vs. LF-ID performance in same settings?
- Real LF better than Random LF?

Compared
- 3-class L1-ID
- 3-class LF-ID (Standalone), real LF
- 3-class LF-ID (Standalone), randomly constructed LF
3-Class L1 vs. 3-Class LF – Results
3-Class L1 vs. 3-Class LF – Results

L1-ID more accurate than LF-ID in same settings

- Basic
- Extended
- Grammar
- InfoGain

L1 (3) vs. LF (3) vs. Random LF (3) vs. Random
3-Class L1 vs. 3-Class LF – Results

Real LF better than random sets of languages

- L1 (3)
- LF (3)
- Random LF (3)
- Random

Categories:
- Basic
- Extended
- Grammar
- InfoGain

Stolerman, Caliskan Islam & Greenstadt
From Language to Family and Back