What similarity tells us about transfer. Retrieving L1 from learner texts in Falko

Marc Reznicek & Felix Golcher

Humboldt-Universität zu Berlin

Zweiter Tübingen-Berlin Workshop zur Analyse von Lernersprache

05.12.2011
Research Questions: Joining two points of view

Transfer

Road map

Our data - the Falko corpus

The similarity measure S – basic concept

Preliminary results

Two issues
  - Taking the essay topic into account
  - Getting rid of copied material

Results

Beyond classification

Conclusion
Research Questions: Joining two points of view

Coming from second language acquisition research

Learner Corpus Research

1. study of learner language
   ▶ patterns
   ▶ controlling variables

What measures can help us uncover hidden patterns in learner data?

Are learner dependent variables detectable in learner texts?

How do those variables affect the learner language?

How strong is the influence of those variables?
Coming from second language acquisition research

Learner Corpus Research
1. study of learner language
   ▶ patterns
   ▶ controlling variables
2. and describe the variability between learners and learner subgroups
Research Questions: Joining two points of view

Coming from second language acquisition research

Learner Corpus Research

1. study of learner language
   ▶ patterns
   ▶ controlling variables

2. and describe the variability between learners and learner subgroups

What measures can help us uncover hidden patterns in learner data?

1. Are learner dependent variables detectable in learner texts?
2. How do those variables affect the learner language?
3. How strong is the influence of those variables?
Coming from stylometry

Stylometry...
Coming from stylometry

Stylometry... 1

classifies texts according to non-linguistic variables:
Research Questions: Joining two points of view

Coming from stylometry

Stylometry...  

1 classifies texts according to non-linguistic variables:  
   ▶ authorship (who wrote a piece?)
Research Questions: Joining two points of view

Coming from stylistometry

Stylistometry...

1. classifies texts according to non-linguistic variables:
   ▶ authorship (who wrote a piece?)
   ▶ gender

Can we apply this technique to learner data?

Can we automatically detect the learners $L_1$ from its texts?

What kind of variables play a (confounding) role?

Can we isolate the influence of different variables?
Research Questions: Joining two points of view

Coming from stylometry

Stylometry...

1. classifies texts according to non-linguistic variables:
   - authorship (who wrote a piece?)
   - gender
   - other such variables.

2. and (ideally) tries to find out the important linguistic features.

Can we apply this technique to learner data?

Can we automatically detect the learners' L1 from its texts?

What kind of variables play a (confounding) role?

Can we isolate the influence of different variables?
Research Questions: Joining two points of view

Coming from stylometry

Stylometry...

1. classifies texts according to non-linguistic variables:
   - authorship (who wrote a piece?)
   - gender
   - other such variables.

2. and (ideally) tries to find out the important linguistic features.
Coming from stylometry

Stylo...m. .y...

1. classifies texts according to non-linguistic variables:
   ▶ authorship (who wrote a piece?)
   ▶ gender
   ▶ other such variables.

2. and (ideally) tries to find out the important linguistic features.

Can we apply this technique to learner data?

1. Can we automatically “detect” the learners $L_1$ from its texts?
2. What kind of variables play a (confounding) role?
3. Can we isolate the influence of different variables?
Converging research questions

1. Can we quantify the influence of the learner’s L1 on his/her language use?
2. How do L1 effects show on different linguistic levels? ▶ lexis ▶ syntax ▶ morphology
3. To what extent do L1 effects lead to ungrammatical structures in the learner language?
4. How strong is the influence of secondary variables (e.g., content)?
Converging research questions

1. Can we quantify the influence of the learner’s L1 on his/her language use?

2. How do L1 effects show on different linguistic levels?
   - lexis
   - syntax
   - morphology
Converging research questions

1. Can we quantifiy the influence of the learner’s L1 on his/her language use?

2. How do L1 effects show on different linguistic levels?
   - lexis
   - syntax
   - morphology

3. To what extent do L1 effects lead to ungrammatical structures in the learner language?
Converging research questions

1. Can we quantify the influence of the learner’s L1 on his/her language use?

2. How do L1 effects show on different linguistic levels?
   - lexis
   - syntax
   - morphology

3. To what extent do L1 effects lead to ungrammatical structures in the learner language?

4. How strong is the influence of secondary variables (e.g. content)?
Research Questions: Joining two points of view

Transfer

Road map

Our data - the Falko corpus

The similarity measure $S$ – basic concept

Preliminary results

Two issues
- Taking the essay *topic* into account
- Getting rid of copied material

Results

Beyond classification

Conclusion
Transfer as cross-linguistic influence

**Transfer - working definition**

Language transfer refers to any instance of learner data where a statistically significant correlation (or probability-based relation) is shown to exist between some feature of the interlanguage and any other language that has been previously acquired (see Ellis 2009).
Transfer as cross-linguistic influence

Transfer - working definition

Language transfer refers to any instance of learner data where a statistically significant correlation (or probability-based relation) is shown to exist between some feature of the interlanguage and any other language that has been previously acquired (see Ellis 2009).

- Many studies have looked at each level independently.
Transfer as cross-linguistic influence

Transfer - working definition

Language transfer refers to any instance of learner data where a statistically significant correlation (or probability-based relation) is shown to exist between some feature of the interlanguage and any other language that has been previously acquired (see Ellis 2009).

- Many studies have looked at each level independently.

Relative contributions of L1 on linguistic levels

[We need] “a reliable way to measure the relative contributions of the native language to the ease or difficulty learners have with each subsystem and, by implication, the total contribution of transfer to the process of second language acquisition.” (Odlin 2003, p. 439)
Research Questions: Joining two points of view

Transfer

Road map

Our data - the Falko corpus

The similarity measure $S$ – basic concept

Preliminary results

Two issues

- Taking the essay *topic* into account
- Getting rid of copied material

Results

Beyond classification

Conclusion
From similarity to transfer

We want to classify IL-texts for author’s L1:

- We define a similarity measure for texts:
  - A text is a string of characters.
  - Take two texts $A$ and $B$, compute a number $S$ from them.
  - Interpret this number as an indicator for similarity.

- Assign a text to the “most similar” L1 (details later!)

*a posteriori* justification

If the assignments are correct,

$\Rightarrow$ then $S$ is a reflection of L1 specific structures in IL (↔ transfer).
From similarity to transfer

Transfer on different linguistic levels

- L1 classification results based on different linguistic levels reflect transfer on that specific level
  - lemma ⇒ (mainly) transfer on lexical choice
  - part-of-speech ⇒ (mainly) syntactic transfer
  - lemma-tok-difference ⇒ inflectional morphology?

If there is a difference between those results for
(a) the learner text
(b) a grammatically corrected version of it (target hypothesis)
then this reflects transfer leading to ungrammatical IL-structures.
From similarity to transfer

Transfer on different linguistic levels
- L1 classification results based on different linguistic levels reflect transfer on that specific level
  - lemma $\Rightarrow$ (mainly) transfer on lexical choice
  - part-of-speech $\Rightarrow$ (mainly) syntactic transfer
  - lemma-tok-difference $\Rightarrow$ inflectional morphology?

Transfer and grammatical errors
- If there is a difference between those results for
  1. the learner text
  2. a grammatically corrected version of it (target hypothesis)
then this reflects transfer leading to ungrammatical IL-structures.
1. Research Questions: Joining two points of view
2. Transfer
3. Road map
4. Our data - the Falko corpus
5. The similarity measure $S$ – basic concept
6. Preliminary results
7. Two issues
   - Taking the essay *topic* into account
   - Getting rid of copied material
8. Results
9. Beyond classification
10. Conclusion
Our data - the Falko corpus

Falko (Lüdeling et al. 2008) corpus subset

**Texts included**
- languages with at least 10 texts
- learners with only one L1

**Very small data sample**
We use only \(\approx 66,000\) tokens. This is 34% of Falko.

<table>
<thead>
<tr>
<th>L1</th>
<th># of texts</th>
</tr>
</thead>
<tbody>
<tr>
<td>German (deu)(^a)</td>
<td>10</td>
</tr>
<tr>
<td>English (eng)</td>
<td>42</td>
</tr>
<tr>
<td>Danish (dan)</td>
<td>37</td>
</tr>
<tr>
<td>French (fra)</td>
<td>14</td>
</tr>
<tr>
<td>Russian (rus)</td>
<td>10</td>
</tr>
<tr>
<td>Turkish (tur)</td>
<td>10</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>126 texts</strong></td>
</tr>
</tbody>
</table>

\(^a\)control group, excluded if sensible

<table>
<thead>
<tr>
<th>title</th>
<th>texts</th>
<th>texts</th>
</tr>
</thead>
<tbody>
<tr>
<td>“crime”</td>
<td>11</td>
<td>Kriminalität zahlt sich nicht aus.</td>
</tr>
<tr>
<td>“feminism”</td>
<td>23</td>
<td>Der Feminismus hat den Interessen der Frauen mehr geschadet als genützt.</td>
</tr>
<tr>
<td>“wages”</td>
<td>60</td>
<td>Die finanzielle Entlohnung eines Menschen sollte dem Beitrag entsprechen, den er/ sie für die Gesellschaft geleistet hat.</td>
</tr>
<tr>
<td>“studies”</td>
<td>32</td>
<td>Die meisten Universitätsabschlüsse sind nicht praxisorientiert und bereiten die Studenten nicht auf die wirkliche Welt vor.</td>
</tr>
</tbody>
</table>
Falko - 6 representations

- We have 6 representations of each text.
- Each representation is defined by two variables:

1. Level of linguistic representation:
   - token: original texts
   - POS: Part-of-Speech tag sequence
   - lemma: lemma sequence

2. Level of error contamination:
   - learner: the raw learner texts
   - Target hypothesis (ZH1): the grammaticalized version (Reznicek et al. 2010)

---

1 Schmid 1994.
Falko - 6 representations

- We have 6 representations of each text.
- Each representation is defined by two variables:
  1. Level of linguistic representation:
     - *token* original texts:
       Man denke an den unterschiedlichen Gruppen, die sich für den Umweltsschutz einsetzen.
     - *POS* Part-of-Speech tag sequence (Treetagger\(^1\)):
       PIS VVFIN APPR ART ADJA NN $, PRELS PRF APPR ART NN VVINF $.
     - *lemma* lemma sequence:
       man denken an d unterschiedlich Gruppe , d er|es|sie für d Umweltsschutz einsetzen .

\(^1\)Schmid 1994.
Falko - 6 representations

- We have 6 representations of each text.
- Each representation is defined by two variables:
  1. Level of linguistic representation:
     - **token** original texts:
       Man denke an den unterschiedlichen Gruppen, die sich für den Umweltsschutz einsetzen.
     - **POS** Part-of-Speech tag sequence (Treetagger$^1$):
       PIS VVFIN APPR ART ADJA NN $,$, PRELS PRF APPR ART NN VVINF $.$
     - **lemma** lemma sequence:
       man denken an d unterschiedlich Gruppe, d er|es|sie für d Umweltsschutz einsetzen .
  2. Level of error contamination:
     - **learner** The raw learner texts:
       Man denke an den unterschiedlichen Gruppen, die [...] 
     - **Target hypothesis (ZH1)**
       the grammaticalized version (Reznicek et al. 2010):
       Man denke an die unterschiedlichen Gruppen, die [...] 

$^1$Schmid 1994.
5 The similarity measure $S$ – basic concept
The similarity measure $S$ – basic concept

$S$ explained by example

Two very short texts:

\[ A = \text{xabay} \]
\[ B = \text{bcbabd} \]

\[ a_2 \log (2 \cdot 1 + 1) = 1.09 \]
\[ a_1 \log (1 \cdot 1 + 1) = 0.69 \]
\[ b_1 \log (1 \cdot 3 + 1) = 1.39 \]
\[ x_1 \log (1 \cdot 0 + 1) = 0.42 \]

\[ S = \sum = 3.17 \]

An important feature

All substrings of all lengths contribute:

⇒ No maximal length is set (as is the usual practice).

No other information than (character) string repetitions are used.

(Humboldt-Universität zu Berlin) Stylometry & transfer in Falko Tübingen II 2011
$S$ explained by example

Two very short texts:

$$A = \text{xabay}$$
**S explained by example**

Two very short texts:

<table>
<thead>
<tr>
<th></th>
<th>A = xabay</th>
<th>B = bcbabd</th>
</tr>
</thead>
</table>

\[ S = \sum \log (2 \cdot 1 + 1) = 1.09 \]

\[ \log (1 \cdot 1 + 1) = 0.69 \]

\[ \log (1 \cdot 3 + 1) = 1.39 \]

\[ \log (1 \cdot 0 + 1) = 0 \]

\[ S = 3.17 \]

An important feature:

- All substrings of all lengths contribute.
- No maximal length is set (as is the usual practice).
- No other information than (character) string repetitions are used.
The similarity measure $S$ – basic concept

**$S$ explained by example**

Two very short texts:

<table>
<thead>
<tr>
<th>substrings</th>
<th>$A = xabay$</th>
<th>$B = bcbabd$</th>
</tr>
</thead>
</table>

$$a_2 \log(2 \cdot 1 + 1) = 1.09$$

$$a_1 \log(1 \cdot 1 + 1) = 0.69$$

$$b_1 \log(1 \cdot 3 + 1) = 1.39$$

$$x_0 \log(1 \cdot 0 + 1) = 0$$

$$S = 3.17$$

An important feature:

All substrings of all lengths contribute:

$\Rightarrow$ No maximal length is set (as is the usual practice).

No other information than (character) string repetitions are used.
**S explained by example**

Two very short texts:

<table>
<thead>
<tr>
<th>substrings</th>
<th>$A = xabay$</th>
<th>$B = bcbabd$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

The similarity measure $S$ – basic concept

All substrings of all lengths contribute:

$S = \sum_{i=1}^{3} \log(2\cdot1+1) = 1.09$,

$S = \sum_{i=1}^{3} \log(1\cdot1+1) = 0.69$,

$S = \sum_{i=1}^{3} \log(1\cdot3+1) = 1.39$,

$S = \sum_{i=1}^{3} \log(1\cdot0+1) = 0.00$.

$S = 3.17$.

An important feature:

No maximal length is set (as is the usual practice).

No other information than (character) string repetitions are used.

(Humboldt-Universität zu Berlin)

Stylometry & transfer in Falko

Tübingen II 2011
S explained by example

Two very short texts:

<table>
<thead>
<tr>
<th>substrings</th>
<th>$A = xab\text{ay}$</th>
<th>$B = bcb\text{abd}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>ab</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

$$S = \sum \log (2 \cdot 1 + 1) = 1.09$$
$$\log (1 \cdot 1 + 1) = 0.69$$
$$\log (1 \cdot 3 + 1) = 1.39$$
$$\log (1 \cdot 0 + 1) = 0$$

All substrings of all lengths contribute:

⇒ No maximal length is set (as is the usual practice).

No other information than (character) string repetitions are used.
The similarity measure $S$ – basic concept

### $S$ explained by example

Two very short texts:

<table>
<thead>
<tr>
<th>substrings</th>
<th>$A = xab\text{ay}$</th>
<th>$B = bcba\text{bd}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>ab</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

$S = \sum \log (2 \cdot 1 + 1) = 1.09$
**S explained by example**

Two very short texts:

<table>
<thead>
<tr>
<th>substrings</th>
<th>A = xabay</th>
<th>B = bcbabbd</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>ab</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>x</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
S explained by example

Two very short texts:

| substrings | $A = xabay$ | $B = bcbabd$ | 2 · 1  
|------------|-------------|---------------|-----------
| a          | 2           | 1             | 2 · 1     |
| ab         | 1           | 1             | 1 · 1     |
| b          | 1           | 3             | 1 · 3     |
| x          | 1           | 0             | 1 · 0     |

$S = \sum$ an important feature

All substrings of all lengths contribute: ⇒ No maximal length is set (as is the usual practice).

No other information than (character) string repetitions are used.
S explained by example

Two very short texts:

<table>
<thead>
<tr>
<th>substrings</th>
<th>A = xabay</th>
<th>B = bcbabd</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>ab</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>x</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
\log(2 \cdot 1) = 0.39 \\
\log(1 \cdot 1) = 0.00 \\
\log(1 \cdot 3) = 1.39 \\
\log(1 \cdot 0) = 0.00
\]

an important feature

All substrings of all lengths contribute:

⇒ No maximal length is set (as is the usual practice).

No other information than (character) string repetitions are used.

(Humboldt-Universität zu Berlin)  
Stylometry & transfer in Falko  
Tübingen II 2011
### S explained by example

Two very short texts:

<table>
<thead>
<tr>
<th>substrings</th>
<th>( A =xabay )</th>
<th>( B = bcbabd )</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>ab</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>x</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
### S explained by example

Two very short texts:

<table>
<thead>
<tr>
<th>substrings</th>
<th>$A = xabay$</th>
<th>$B = bcbabd$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>ab</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>x</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

$S$ is an important feature:
- All substrings of all lengths contribute.
- There is no maximal length set (as is usual practice).
- No other information than (character) string repetitions are used.
The similarity measure $S$ – basic concept

$S$ explained by example

Two very short texts:

<table>
<thead>
<tr>
<th>substrings</th>
<th>$A = xabay$</th>
<th>$B = bcbabd$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>ab</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>x</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The similarity measure $S$ – basic concept

$S$ explained by example

Two very short texts:

<table>
<thead>
<tr>
<th>substrings</th>
<th>$A = xabaya$</th>
<th>$B = bcbabd$</th>
<th>$\log(2 \cdot 1 + 1) = 1.09$</th>
<th>$\log(1 \cdot 1 + 1) = 0.69$</th>
<th>$\log(1 \cdot 3 + 1) = 1.39$</th>
<th>$\log(1 \cdot 0 + 1) = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ab</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$$S = \sum = 3.17$$

An important feature:
- All substrings of all lengths contribute.
- No maximal length is set (as is the usual practice).
- No other information than (character) string repetitions are used.
The similarity measure $S$ – basic concept

$S$ explained by example

Two very short texts:

<table>
<thead>
<tr>
<th>substrings</th>
<th>$A = xabay$</th>
<th>$B = bcbabd$</th>
<th>subtrange lengths</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2</td>
<td>1</td>
<td>$\log(2 \cdot 1 + 1) = 1.09$</td>
</tr>
<tr>
<td>ab</td>
<td>1</td>
<td>1</td>
<td>$\log(1 \cdot 1 + 1) = 0.69$</td>
</tr>
<tr>
<td>b</td>
<td>1</td>
<td>3</td>
<td>$\log(1 \cdot 3 + 1) = 1.39$</td>
</tr>
<tr>
<td>x</td>
<td>1</td>
<td>0</td>
<td>$\log(1 \cdot 0 + 1) = 0$</td>
</tr>
</tbody>
</table>

$S = \sum = 3.17$

an important feature

All substrings of all lengths contribute:

$\Rightarrow$ No maximal length is set (as is the usual praxis).

No other information than (character) string repetitions are used.
Various stylometric tasks have been investigated with $S$:


Some details of the classification method

- Take one text $T_i$ after another as test text (126 texts).
- following steps:
  1. Compute $S(T_i, T_j)$ for the remaining 125 training texts ($i \neq j$)
  2. Group those $S$ values according to the L1 of those training texts.
  3. Compute the mean $S$ value $S_{L1}$ for each L1 group.
  4. Assign the test text $T_i$ to the L1 group with the highest $S_{L1}$.

(Humboldt-Universität zu Berlin) Stylometry & transfer in Falko Tüberlin II 2011 17 / 52
Research Questions: Joining two points of view

Transfer

Road map

Our data - the Falko corpus

The similarity measure $S$ – basic concept

Preliminary results

Two issues
- Taking the essay topic into account
- Getting rid of copied material

Results

Beyond classification

Conclusion
L1 classification – Proof of concept

Sensible first results:
- Above baseline (31.2 = 28%).
  - Reproducing similar results\(^a\).
- *tok* and *lemma* nearly identical.
- *POS* lower.
- *Target Hypothesis* seems lower.

\(^a\)Koppel et al. 2003; Koppel et al. 2005; Tsur et al. 2009; Golcher to appear

**Figure**: disregarding German L1 texts.
1 Research Questions: Joining two points of view
2 Transfer
3 Road map
4 Our data - the Falko corpus
5 The similarity measure $S$ – basic concept
6 Preliminary results

7 Two issues
   - Taking the essay *topic* into account
   - Getting rid of copied material

8 Results
9 Beyond classification
10 Conclusion
Research Questions: Joining two points of view

Transfer

Road map

Our data - the Falko corpus

The similarity measure $S$ – basic concept

Preliminary results

Two issues
  - Taking the essay *topic* into account
  - Getting rid of copied material

Results

Beyond classification

Conclusion
Another possible influence: Content

- Until now we ignored the \textit{essay topic} people wrote about.
- Obviously, texts about “crime” will share words.
- This of course leads to higher $S$ values.
- If this \textit{topic} effect is larger than the L1 effect, the latter will be masked.

In stylometry, this is a well known problem.
classification according to \textit{topic}

- \textit{tok} and \textit{lemma} very high (> 98%).
- \textit{POS} much lower.
  - but above baseline!

\textbf{\Rightarrow} \textit{topic} effect very strong.
A simple heuristic for filtering out essay topic

- We divide all $S(A, B)$ in two groups:
  1. $A$ and $B$ have the same topic.
  2. They have not.
- We compute the mean of each group.
- Each $S$ value is divided by the mean of its group.
1. Research Questions: Joining two points of view
2. Transfer
3. Road map
4. Our data - the Falko corpus
5. The similarity measure $S$ – basic concept
6. Preliminary results
7. Two issues
   - Taking the essay *topic* into account
   - Getting rid of copied material
8. Results
9. Beyond classification
10. Conclusion
Copied material

Texts about the same subject will normally share lexical material. We have an additional problem:

- The full topic we call “feminism” reads as
  
  \[\text{Der Feminismus hat den Interessen der Frauen mehr geschadet als genützt.}\]
  
  Feminism damaged the interests of the women rather than it helped them.

- Especially learners tend to copy phrases like “\text{den Interessen der Frauen}”.

- These long shared substrings make unproportional contributions to $S$. 

explosion of substrings

The number of substrings of a string grows quadratically with its length.
Definition of “copied material”

We use a simple heuristic to identify copied material

Definition (copied material of order $n$)

A string in text $T$ is copied from source text $S$, if

- it occurs only once in the source text $S$.
- this is true even if we strip $n$ characters at both sides.

Example (set $n$ to 1)

source $S$  Do we have beer or do we have wine, Josef?

text $T$  Someone must have been telling lies about Josef K.

applying the definition:

“Josef” is copied.

“have b” is not (“have” occurs twice in source text $S$)
Definition of “copied material”

We use a simple heuristic to identify copied material

**Definition (copied material of order $n$)**

A string in text $T$ is copied from source text $S$, if

... it occurs only once in the source text $S$.

... this is true even if we strip $n$ characters at both sides.

**Example (set $n$ to 1)**

source $S$  Do we have beer or do we have wine, Josef?

text $T$  Someone must have been telling lies about Josef K.

applying the definition:

“Josef” is copied.

“have b” is not (“have” occurs twice in source text $S$)
Definition of “copied material”

We use a simple heuristic to identify copied material

**Definition (copied material of order $n$)**

A string in text $T$ is copied from source text $S$, if

- it occurs only once in the source text $S$.
- this is true even if we strip $n$ characters at both sides.

**Example (set $n$ to 1)**

| source $S$ | Do we **have** beer or do we have wine, Josef? |
| text $T$   | Someone must **have been** telling lies about Josef K. |

applying the definition:

- “Josef” is copied.
- “have b” is not (“have” occurs twice in source text $S$)
Two issues

Getting rid of copied material

Definition of “copied material”

We use a simple heuristic to identify copied material

Definition (copied material of order $n$)

A string in text $T$ is copied from source text $S$, if

- it occurs only once in the source text $S$.
- this is true even if we strip $n$ characters at both sides.

Example (set $n$ to 1)

source $S$  Do we have beer or do we have wine, Josef?

text $T$  Someone must have been telling lies about Josef K.

applying the definition:

“Josef” is copied.

“have b” is not (“have” occurs twice in source text $S$)
Two issues

Example

\[ n = 2 \]

Zum Schluss glaube ich, dass der Feminismus den Interessen der Frauen sehr viel nützen könne, aber es gibt zu viele Leute, die die Konzepte des Feminismus schaden, wenn sie dem Feminismus für falschen Gründen oder in den falschen Situationen nützen.

At the end I think, that feminism could help the interests of the women very much, but there are too many people, which harm them concepts of feminism, if they help femininism for wrongs reasons or in wrong situations.
Example

\[ n = 5 \]

Zum Schluss glaube ich, dass der Feminismus den Interessen der Frauen sehr viel nützen könne, aber es gibt zu viele Leute, die die Konzepte des Feminismus schaden, wenn sie dem Feminismus für falschen Gründen oder in den falschen Situationen nützen.

At the end I think, that feminism could help the interests of the women very much, but there are too many people, which harm them concepts of feminism, if they help feminism for wrongs reasons or in wrong situations.
Two issues

Getting rid of copied material

Example

\[ n = 10 \]

Zum Schluss glaube ich, dass der Feminismus den Interessen der Frauen sehr viel nützen könne, aber es gibt zu viele Leute, die die Konzepte des Feminismus schaden, wenn sie dem Feminismus für falschen Gründen oder in den falschen Situationen nützen.

At the end I think, that feminism could help the interests of the women very much, but there are too many people, which harm them concepts of feminism, if they help femininism for wrongs reasons or in wrong situations.
Research Questions: Joining two points of view

Transfer

Road map

Our data - the Falko corpus

The similarity measure $S$ – basic concept

Preliminary results

Two issues
- Taking the essay *topic* into account
- Getting rid of copied material

Results

Beyond classification

Conclusion
Results for $L1$ classification

Classification Quality

Figure: Classified by $L1$. Without German $L1$. Horizontal line is base line.
Results for \(L1\) classification

**Figure:** Classified by \(L1\). Without German \(L1\). Horizontal line is base line.

**observation 1**

Filtering out copied material helps a lot for \(tok\) and \(lemma\).

\[\Rightarrow\] Copied material hampers \(L1\) classification.

Optimum between \(n = 5\) and \(n = 10\).
Results for L1 classification

**Figure:** Classified by L1. Without German L1. Horizontal line is base line.
Results for L1 classification

Figure: Classified by L1. Without German L1. Horizontal line is base line.

observation 2

Filtering out copied material does not change much for POS
Results for $L1$ classification

Figure: Classified by $L1$. Without German $L1$. Horizontal line is base line.
Results for L1 classification

![Graph showing results for L1 classification.](image)

**Figure:** Classified by L1. Without German L1. Horizontal line is base line.

**observation 3**

\[ \text{tok} > \text{lemma} \gg \text{POS} \]

**POS** is a very reduced text version.
Results for L1 classification

Figure: Classified by L1. Without German L1. Horizontal line is base line.
Results for L1 classification

Figure: Classified by L1. Without German L1. Horizontal line is base line.

observation 4

To average out topic helps for tok and lemma.
⇒ Ignoring topic hampers L1 classification.
Results for $L1$ classification

![Graph](image)

Figure: Classified by $L1$. Without German $L1$. Horizontal line is base line.
Results for $L1$ classification

Figure: Classified by $L1$. Without German $L1$. Horizontal line is base line.

observation 5

Again, no such effect for $POS$.  
$\Rightarrow$ much less interaction between $topic$ and $L1$.  

**Results for L1 classification**

Figure: Classified by L1. Without German L1. Horizontal line is base line.
Results for L1 classification

**Figure:** Classified by L1. Without German L1. Horizontal line is base line.

**observation 6**

learner text > ZH1

Correction reduces L1 effect. Not so clear for lemma.
Results for *topic* classification

![Graph showing results for topic classification](image)

**Figure:** Classified by text *topic*. With German files. Horizontal line is base line.
Results for *topic* classification

**Figure:** Classified by text *topic*. With German files. Horizontal line is base line.
Results for *topic* classification

**Figure:** Classified by text *topic*. With German files. Horizontal line is base line.

**Observation 1**

Filtering out *copied material* with high $n$ does not influence *tok* and *lemma*.

$\Rightarrow$ *Copied material* is not identical with *text topic*. 
Results for topic classification

Figure: Classified by text topic. With German files. Horizontal line is base line.
Results for topic classification

![Graph showing topic classification results.]

**Figure:** Classified by text topic. With German files. Horizontal line is base line.

**Observation 2**

Again, between $n = 5$ and $n = 10$ is the most interesting stretch.
Results for *topic* classification

**Figure:** Classified by text *topic*. With German files. Horizontal line is base line.
Results for *topic* classification

**Figure:** Classified by text *topic*. With German files. Horizontal line is base line.

**observation 3**

When more and more is filtered out nearly hits the base line. ⇒ Much of L1 influence in *topic* due to copied material. ALL?
Results for *topic* classification

Figure: Classified by text *topic*. With German files. Horizontal line is base line.

**observation 3**

When more and more is filtered out **POS** nearly hits the base line.
⇒ Much of L1 influence in *topic* due to *copied material*. **ALL?**
Results for *topic* classification

![Graph showing results for topic classification](image)

**Figure**: Classified by text *topic*. With German files. Horizontal line is base line.
Results for *topic* classification

**Figure**: Classified by text *topic*. With German files. Horizontal line is base line.

**Observation 4**

\[
\text{lemma} > \text{tok} \quad (\Rightarrow \text{POS})
\]

\(\Rightarrow\) *lemma* better for *topic*, *tok* better for L1 classification.
Results for *topic* classification

![Graphs showing results for topic classification.](image)

**Figure:** Classified by text *topic*. With German files. Horizontal line is base line.
Results for *topic* classification

**Figure:** Classified by text *topic*. With German files. Horizontal line is base line.

**observation 5**

Our heuristic for *topic* influence reduction works very well: Performance drops to estimated base line.
<table>
<thead>
<tr>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>
Where to go from here?

- Successful classification is a reliable indicator for existing transfer.
  but effect sizes can’t be readily quantified.
- The *topic* effect seems to be “stronger” than L1.
  but how much?
  ⇒ comparison of classification accuracies is rather indirect.

Can we surpass the *stylometric* classificational view?

1. Can we directly quantify the influence of *topic* and L1?
2. Can we directly compare them? For different levels of representation?
Building a (linear mixed) model

- For each $S(A, B)$ we construct two variables:
  - `sameTopic` 1 if $A$ and $B$ share its topic, 0 otherwise.
  - `sameL1` 1 if authors of $A$ and $B$ share L1, 0 otherwise.

- Now we set up a model

$$S = \alpha \cdot \text{sameTopic} + \beta \cdot \text{sameL1} + \text{<text specific contributions>} + \epsilon$$

where
- $\epsilon$ is a normally distributed error term.
- the `<text specific contributions>` are assumed normally distributed too.
Building a (linear mixed) model

- For each $S(A, B)$ we construct two variables:
  
  **sameTopic** 1 if $A$ and $B$ share its topic, 0 otherwise.
  
  **sameL1** 1 if authors of $A$ and $B$ share L1, 0 otherwise.

- Now we set up a model

  $$S = \alpha \cdot \text{sameTopic} + \beta \cdot \text{sameL1} + \langle \text{text specific contributions} \rangle + \epsilon$$

  where

  - $\epsilon$ is a normally distributed error term.
  - the $\langle \text{text specific contributions} \rangle$ are assumed normally distributed too.
Building a (linear mixed) model

- For each $S(A, B)$ we construct two variables:
  - \texttt{sameTopic} 1 if $A$ and $B$ share its topic, 0 otherwise.
  - \texttt{sameL1} 1 if authors of $A$ and $B$ share L1, 0 otherwise.
- Now we set up a model

$$S = \alpha \cdot \texttt{sameTopic} + \beta \cdot \texttt{sameL1} + \langle\text{text specific contributions}\rangle + \epsilon$$

where
- $\epsilon$ is a normally distributed error term.
- the $\langle\text{text specific contributions}\rangle$ are assumed normally distributed too.
Building a (linear mixed) model

- For each $S(A, B)$ we construct two variables:
  - `sameTopic` 1 if $A$ and $B$ share its topic, 0 otherwise.
  - `sameL1` 1 if authors of $A$ and $B$ share L1, 0 otherwise.

- Now we set up a model

\[ S = \alpha \cdot \text{sameTopic} + \beta \cdot \text{sameL1} + \text{<text specific contributions>} + \epsilon \]

where
- $\epsilon$ is a normally distributed error term.
- the `<text specific contributions>` are assumed normally distributed too.
Building a (linear mixed) model

- For each $S(A, B)$ we construct two variables:
  - `sameTopic` 1 if $A$ and $B$ share its topic, 0 otherwise.
  - `sameL1` 1 if authors of $A$ and $B$ share L1, 0 otherwise.
- Now we set up a model

\[ S = \alpha \cdot \text{sameTopic} + \beta \cdot \text{sameL1} + \langle \text{text specific contributions} \rangle + \epsilon \]

where

- $\epsilon$ is a normally distributed error term.
- the $\langle \text{text specific contributions} \rangle$ are assumed normally distributed too.
Building a (linear mixed) model

- For each $S(A, B)$ we construct two variables:
  
  \[
  \text{sameTopic} \quad \begin{cases} 
  1 & \text{if } A \text{ and } B \text{ share its topic,} \\
  0 & \text{otherwise.} 
  \end{cases}
  \]

  \[
  \text{sameL1} \quad \begin{cases} 
  1 & \text{if authors of } A \text{ and } B \text{ share L1,} \\
  0 & \text{otherwise.} 
  \end{cases}
  \]

- Now we set up a model

\[
S = \alpha \cdot \text{sameTopic} + \beta \cdot \text{sameL1} + \langle \text{text specific contributions} \rangle + \epsilon
\]

where

- $\epsilon$ is a normally distributed error term.
- the $\langle \text{text specific contributions} \rangle$ are assumed normally distributed too.

- This (linear mixed) model is fitted.
Building a (linear mixed) model

- For each $S(A, B)$ we construct two variables:
  - $\text{sameTopic} = 1$ if $A$ and $B$ share its topic, 0 otherwise.
  - $\text{sameL1} = 1$ if authors of $A$ and $B$ share L1, 0 otherwise.
- Now we set up a model

$$S = \alpha \cdot \text{sameTopic} + \beta \cdot \text{sameL1} + \text{<text specific contributions>} + \epsilon$$

where

- $\epsilon$ is a normally distributed error term.
- the $\text{<text specific contributions>}$ are assumed normally distributed too.

- This (linear mixed) model is fitted.
- The parameters $\alpha$ and $\beta$ can be compared.
The results

Figure: L1 ($\beta$) effect divided by topic ($\alpha$) effect.

observations

1. essay topic very strong.
Beyond classification

The results

Figure: L1 ($\beta$) effect divided by topic ($\alpha$) effect.

observations

1. essay topic very strong.
2. Much stronger than L1 for token and lemma.
The results

Figure: L1 ($\beta$) effect divided by topic ($\alpha$) effect.

observations

1. essay topic very strong.
2. Much stronger than L1 for token and lemma.
3. No difference between token and lemma.
Beyond classification

The results

Figure: L1 ($\beta$) effect divided by topic ($\alpha$) effect.

observations

1. essay topic very strong.
2. Much stronger than L1 for token and lemma.
3. No difference between token and lemma.
4. the L1 influence in POS is much more pronounced.

(Humboldt-Universität zu Berlin)  Stylometry & transfer in Falko  Tüberlin II 2011  35 / 52
Beyond classification

The results

![Diagram showing the relative influence of L1 and topic learner]

**Observations**

1. Essay topic very strong.
2. Much stronger than L1 for token and lemma.
3. No difference between token and lemma.
4. The L1 influence in POS is much more pronounced.
5. Removing errors (slightly) weakens L1 influence.

**Figure:** L1 ($\beta$) effect divided by topic ($\alpha$) effect.
Research Questions: Joining two points of view

Transfer

Road map

Our data - the Falko corpus

The similarity measure $S$ – basic concept

Preliminary results

Two issues
  - Taking the essay *topic* into account
  - Getting rid of copied material

Results

Beyond classification

Conclusion
Main results

1. $L_1$ has a high and quantifiable influence on our similarity measure $S$. 
Main results

1. $L_1$ has a high and quantifiable influence on our similarity measure $S$.
2. Sensible relations between learner text and Target Hypothesis.
Main results

1. \( L1 \) has a high and quantifiable influence on our similarity measure \( S \).
2. Sensible relations between learner text and \( \text{Target Hypothesis} \).
3. Sensible relations between \( \text{tok, lemma} \) and \( \text{POS} \).
Main results

1. $L_1$ has a high and quantifiable influence on our similarity measure $S$.
2. Sensible relations between learner text and Target Hypothesis.
3. Sensible relations between tok, lemma and POS.
4. We see a vivid interplay between $L_1$ and topic.
**topic**

Not to take topic into account might ignore a strong source of variance.

**copied material**

Copied material can distort results.

⇒ This can and should be handled independently.
Todos

Looking deeper
Which features of the learner texts contain the L1 dependence?

POS and topic
How strong is the influence of text topic on the POS representation? Is it spurious or linguistically interesting?
Thank you


Literatur III


Another view on the results

<table>
<thead>
<tr>
<th>Level of Representation</th>
<th>Correctly Classified Files (L1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>80</td>
</tr>
</tbody>
</table>

Figure: L1 Classification. Maximum at $82/113 = 0.72 \pm 0.09$. 
Distribution of right and wrong classifications

<table>
<thead>
<tr>
<th>classified as</th>
<th>dan</th>
<th>deu</th>
<th>eng</th>
<th>fra</th>
<th>rus</th>
<th>tur</th>
</tr>
</thead>
<tbody>
<tr>
<td>real L1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dan</td>
<td>19</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>9</td>
<td>37</td>
</tr>
<tr>
<td>deu</td>
<td>12</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>13</td>
</tr>
<tr>
<td>eng</td>
<td>2</td>
<td>6</td>
<td>11</td>
<td>8</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>fra</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>8</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>rus</td>
<td></td>
<td>7</td>
<td></td>
<td>2</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>tur</td>
<td></td>
<td></td>
<td>2</td>
<td>8</td>
<td></td>
<td>10</td>
</tr>
</tbody>
</table>

Figure: Raw text.
Distribution of right and wrong classifications

**Figure:** title averaged out.
Distribution of right and wrong classifications

![Distribution of Right and Wrong Classifications](image)

**Figure**: copied material removed.
Distribution of right and wrong classifications

<table>
<thead>
<tr>
<th>real L1</th>
<th>dan</th>
<th>deu</th>
<th>eng</th>
<th>fra</th>
<th>rus</th>
<th>tur</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. **German** is detected with 100% accuracy.

Figure: copied material removed.
Distribution of right and wrong classifications

1. **German** is detected with 100% accuracy.
   - IL has been claimed to be more variable. (see Romaine 2003)

Figure: copied material removed.
Distribution of right and wrong classifications

1. **German** is detected with 100% accuracy.
   - IL has been claimed to be more variable.
     (see Romaine 2003)

2. Most classification errors occur for **English** learners.

---

**Figure:** copied material removed.
Distribution of right and wrong classifications

<table>
<thead>
<tr>
<th>real L1</th>
<th>classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>dan</td>
<td>dan</td>
</tr>
<tr>
<td>deu</td>
<td>deu</td>
</tr>
<tr>
<td>eng</td>
<td>eng</td>
</tr>
<tr>
<td>fra</td>
<td>fra</td>
</tr>
<tr>
<td>rus</td>
<td>rus</td>
</tr>
<tr>
<td>tur</td>
<td>tur</td>
</tr>
</tbody>
</table>

1. **German** is detected with 100% accuracy.
   - IL has been claimed to be more variable.
     (see Romaine 2003)

2. Most classification errors occur for **English** learners.
   - Influence of common English L2 on German L3?
     (see Cook 2003)

**Figure:** copied material removed.
Distribution of right and wrong classifications

1. **German** is detected with 100% accuracy.
   - IL has been claimed to be more variable. (see Romaine 2003)

2. Most classification errors occur for **English** learners.
   - Influence of common English L2 on German L3? (see Cook 2003)

3. **Turkish** behaves a bit erratic.

**Figure:** copied material removed.
Distribution of right and wrong classifications

1. **German** is detected with 100% accuracy.
   - IL has been claimed to be more variable. (see Romaine 2003)
2. Most classification errors occur for **English** learners.
   - Influence of common English L2 on German L3? (see Cook 2003)
3. **Turkish** behaves a bit erratic.
   - Those were the most ungrammatical texts.

Figure: copied material removed.
Distribution of right and wrong classifications

<table>
<thead>
<tr>
<th>real L1</th>
<th>classified as</th>
<th>dan</th>
<th>deu</th>
<th>eng</th>
<th>fra</th>
<th>rus</th>
<th>tur</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>19</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>9</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>6</td>
<td>11</td>
<td>8</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>8</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>7</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>10</td>
</tr>
</tbody>
</table>

![Figure](image)

*Figure: Raw text.*
Distribution of right and wrong classifications

Figure: title averaged out.
### Distribution of right and wrong classifications

<table>
<thead>
<tr>
<th>real L1</th>
<th>classified as</th>
<th>dan</th>
<th>deu</th>
<th>eng</th>
<th>fra</th>
<th>rus</th>
<th>tur</th>
</tr>
</thead>
<tbody>
<tr>
<td>dan</td>
<td></td>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>deu</td>
<td></td>
<td></td>
<td>5</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eng</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fra</td>
<td></td>
<td></td>
<td></td>
<td>13</td>
<td>4</td>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td>rus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>tur</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>42</td>
</tr>
</tbody>
</table>

|          |               | 37  | 13  | 42  | 14  | 10  | 10  |
|          |               | 23  | 22  | 29  | 15  | 16  | 21  |

**Figure**: copied material removed.
Distribution of right and wrong classifications

**Figure:** copied material removed.

1. **German** is detected with 100% accuracy.
Distribution of right and wrong classifications

1. **German** is detected with 100% accuracy.
   - IL has been claimed to be more variable. (see Romaine 2003)

Figure: copied material removed.
Distribution of right and wrong classifications

1. **German** is detected with 100% accuracy.
   - IL has been claimed to be more variable. (see Romaine 2003)
2. Most classification errors occur for **English** learners.

Figure: copied material removed.
Distribution of right and wrong classifications

1. **German** is detected with 100% accuracy.
   - IL has been claimed to be more variable. (see Romaine 2003)

2. Most classification errors occur for **English** learners.
   - Influence of common English L2 on German L3? (see Cook 2003)

**Figure:** copied material removed.
Distribution of right and wrong classifications

1. **German** is detected with 100% accuracy.
   - IL has been claimed to be more variable. (see Romaine 2003)

2. Most classification errors occur for **English** learners.
   - Influence of common English L2 on German L3? (see Cook 2003)

3. **Turkish** behaves a bit erratic.

**Figure:** copied material removed.

---

<table>
<thead>
<tr>
<th>real L1</th>
<th>classified as</th>
<th>dan</th>
<th>deu</th>
<th>eng</th>
<th>fra</th>
<th>rus</th>
<th>tur</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>22</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>21</td>
<td>4</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>23</td>
<td>22</td>
<td>29</td>
<td>15</td>
<td>16</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>126</td>
</tr>
</tbody>
</table>

(Humboldt-Universität zu Berlin)  Stylometry & transfer in Falko  Tüberlin II 2011  47 / 52
Distribution of right and wrong classifications

1. **German** is detected with 100% accuracy.
   - IL has been claimed to be more variable. (see Romaine 2003)

2. Most classification errors occur for **English** learners.
   - Influence of common English L2 on German L3? (see Cook 2003)

3. **Turkish** behaves a bit erratic.
   - Those were the most ungrammatical texts.

*Figure:* copied material removed.
Norming S

Density plots
An obvious problem

The similarity measure $S$ as a formula

$$S(A, B) = \sum_{\text{all substrings } s} \log(F_A(s)F_B(s) + 1)$$

$F_A(s)$ – Frequency of substring $s$ in Text $A$

- Longer texts $\Rightarrow$ more and more frequent substrings.
- $S$ grows with text length!
- Length dependency not easy to parametrize.
- and that would not be the full story...
- An working heuristic is applied.
A life example

- Eight Dutch authors\(^a\).
- One training file / one test file.
- Each training file compared with each test file.

$\Rightarrow$ Training File 8 is the shortest one.

$\Rightarrow$ Darkest column.

$\Rightarrow$ lowest $S$ values.

\(^a\)Juola 2004.

**Figure:** Dark: low $S$-values; Light: high $S$-values.

**Simple:** Dividing Columns by their mean.
Averaging out single text dependencies

This normed version of $S$ is what we really used.
Norming $S$

Density plots
Distribution of $S(A, B)$ values

**Green:** $A$ and $B$ share title or L1

**Red:** Different title or L1.

**Same title or not?**

- **Green:** Much stronger than L1.
- **Red:** But similarity due to L1 is what we are interested in.
Distribution of $S(A, B)$ values after averaging out title

Again: Green: $A$ and $B$ share L1; Red: Different L1.

with title:

- The difference is much clearer now.
- Classification jumps from 65 to 74 correct decisions (out of 126).
Distribution of $S(A, B)$ values after averaging out title
Again: Green: $A$ and $B$ share $L1$; Red: Different $L1$.

The difference is much clearer now.
Classification jumps from 65 to 74 correct decisions (out of 126).
Suspiciously stretched right tail.
Distribution of $S(A, B)$ values after averaging out title

Again: **Green**: $A$ and $B$ share L1; **Red**: Different L1.

with *title*:

- The difference is much clearer now.
- Classification jumps from 65 to 74 correct decisions (out of 126).
- Suspiciously stretched right tail. $\Rightarrow$ To this we turn now.
The right tail is greatly reduced.

Classification results again jump from 74 to 84 correct (from 126).