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

Coming from second language acquisition research

Learner Corpus Research

1. study of learner language
   - patterns
   - controlling 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

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:

Coming from stylometry

Stylometry. . .

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

Coming from stylometry

Stylometry. . .

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

Coming from stylometry

Stylometry...

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

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

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?

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?

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 quantify the influence of the learner's L1 on his/her language use?
2. How do *L*1 effects show on different linguistic levels?
   - lexis
   - syntax
   - morphology
3. To what extent do *L*1 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 *L*1 effects show on different linguistic levels?
    - lexis
    - syntax
    - morphology
3. To what extent do *L*1 effects lead to ungrammatical structures in the learner language?
4. How strong is the influence of secondary variables (e.g. content)?

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)

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.
  - Interprete 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 ($\Leftarrow$ 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?

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?

Transfer and grammatical errors

- 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.

Falko (Lüdeling et al. 2008) corpus subset

Falko

Texts included

- languages with at least 10 texts
- learners with only one L1

Very small data sample

We use only ≈ 66.000 tokens.
This is 34% of Falko.

L1		# of texts
German	(deu)[a]	10
English	(eng)	42
Danish	(dan)	37
French	(fra)	14
Russian	(rus)	10
Turkish	(tur)	10
total		**126 texts**

[a] control group, excluded if sensible

title	texts	
"crime"	11	Kriminalität zahlt sich nicht aus.
"feminism"	23	Der Feminismus hat den Interessen der Frauen mehr geschadet als genützt.
"wages"	60	Die finanzielle Entlohnung eines Menschen sollte dem Beitrag entsprechen, den er/ sie für die Gesellschaft geleistet hat.
"studies"	32	Die meisten Universitätsabschlüsse sind nicht praxisorientiert und bereiten die Studenten nicht auf die wirkliche Welt vor.

Falko - 6 representations

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

[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.

S explained by example

Two very short texts:

S explained by example

Two very short texts:

	$A = \texttt{xabay}$		

S explained by example

Two very short texts:

	$A = \texttt{xabay}$	$B = \texttt{bcbabd}$	

S explained by example

Two very short texts:

substrings	$A = \texttt{xabay}$	$B = \texttt{bcbabd}$	

S explained by example

Two very short texts:

substrings	$A =$ `xabay`	$B =$ `bcbabd`	
a	2	1	

S explained by example

Two very short texts:

substrings	$A =$ xabay	$B =$ bcbabd	
a	2	1	
ab	1	1	

S explained by example

Two very short texts:

substrings	$A =$ xabay	$B =$ bcbabd	
a	2	1	
ab	1	1	
b	1	3	

S explained by example

Two very short texts:

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

S explained by example

Two very short texts:

substrings	$A = \mathtt{xabay}$	$B = \mathtt{bcbabd}$	
a	2	1	$2 \cdot 1$
ab	1	1	$1 \cdot 1$
b	1	3	$1 \cdot 3$
x	1	0	$1 \cdot 0$

S explained by example

Two very short texts:

substrings	$A = \texttt{xabay}$	$B = \texttt{bcbabd}$	
a	2	1	$\log(2 \cdot 1 \quad)$
ab	1	1	$\log(1 \cdot 1 \quad)$
b	1	3	$\log(1 \cdot 3 \quad)$
x	1	0	$\log(1 \cdot 0 \quad)$

S explained by example

Two very short texts:

substrings	$A = \texttt{xabay}$	$B = \texttt{bcbabd}$	
a	2	1	$\log(2 \cdot 1 + 1)$
ab	1	1	$\log(1 \cdot 1 + 1)$
b	1	3	$\log(1 \cdot 3 + 1)$
x	1	0	$\log(1 \cdot 0 + 1)$

S explained by example

Two very short texts:

substrings	$A =$ xabay	$B =$ bcbabd	
a	2	1	$\log(2 \cdot 1 + 1) = 1.09$
ab	1	1	$\log(1 \cdot 1 + 1) = 0.69$
b	1	3	$\log(1 \cdot 3 + 1) = 1.39$
x	1	0	$\log(1 \cdot 0 + 1) = 0$

S explained by example

Two very short texts:

substrings	$A = \texttt{xabay}$	$B = \texttt{bcbabd}$	
a	2	1	$\log(2 \cdot 1 + 1) = 1.09$
ab	1	1	$\log(1 \cdot 1 + 1) = 0.69$
b	1	3	$\log(1 \cdot 3 + 1) = 1.39$
x	1	0	$\log(1 \cdot 0 + 1) = 0$
			$\sum = 3.17$

S explained by example

Two very short texts:

substrings	$A = \mathtt{xabay}$	$B = \mathtt{bcbabd}$	
a	2	1	$\log(2 \cdot 1 + 1) = 1.09$
ab	1	1	$\log(1 \cdot 1 + 1) = 0.69$
b	1	3	$\log(1 \cdot 3 + 1) = 1.39$
x	1	0	$\log(1 \cdot 0 + 1) = 0$
			$S = \sum = 3.17$

S explained by example

Two very short texts:

substrings	$A =$ xabay	$B =$ bcbabd	
a	2	1	$\log(2 \cdot 1 + 1) = 1.09$
ab	1	1	$\log(1 \cdot 1 + 1) = 0.69$
b	1	3	$\log(1 \cdot 3 + 1) = 1.39$
x	1	0	$\log(1 \cdot 0 + 1) = 0$
			$S = \sum = 3.17$

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:

Felix Golcher (2007). “A new text statistical measure and its application to stylometry”. In: *Corpus Linguistics 2007*. University of Birmingham

Felix Golcher (to appear). “Analysing counting suffix trees of natural language texts (preliminary title)”. PhD thesis. Humboldt-Universität zu Berlin

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 $\overline{S}_{L1}$ for each L1 group.
  4. Assign the test text T_i to the L1 group with the highest $\overline{S}_{L1}$.

L1 classification – Proof of concept

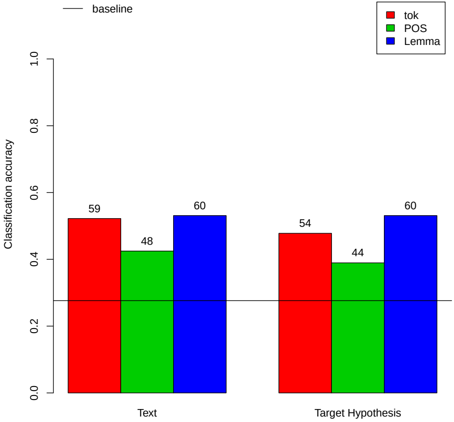


Figure: disregarding German L1 texts.

Sensible first results:

- Above baseline ($31.2 = 28\%$).
  - Reproducing similar results[a].
- *tok* and *lemma* nearly identical.
- *POS* lower.
- *Target Hypothesis* seems lower.

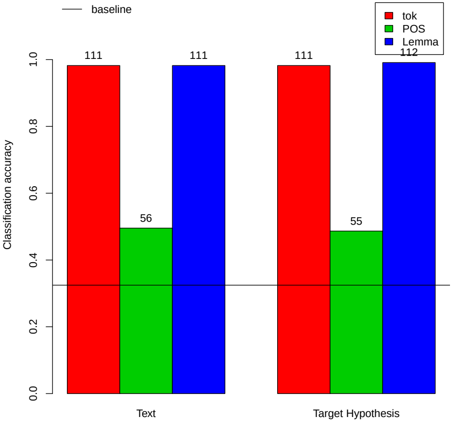
[a] Koppel et al. 2003; Koppel et al. 2005; Tsur e 2009; Golcher to appear

Another possible influence: Content

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

In stylometry, this is a well known problem.

classification according to *topic*



- *tok* and *lemma* very high ($>$ 98%).
- *POS* much lower.
  - but above baseline!

=> *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.

Copied material

explosion of substrings

The number of substrings of a string grows quadratically with its length.

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

- The full topic we call "feminism" reads as

  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 "den Interessen der Frauen".
- These long shared substrings make unproportional contributions to 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)

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)

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 Feminisumus 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 Feminisumus 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

$$n = 10$$

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 Feminisumus 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.

Results for $L1$ classification

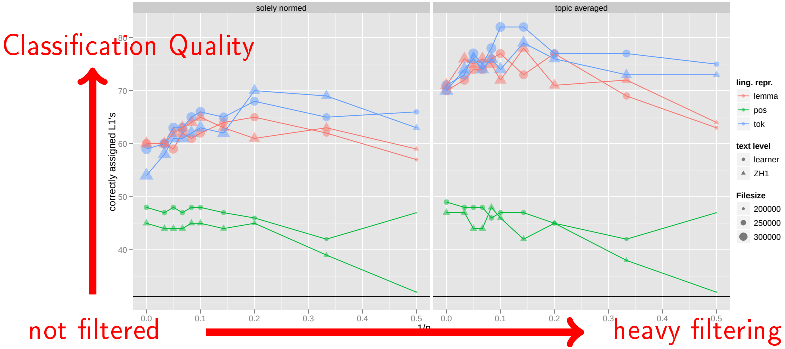


Figure: Classified by $L1$. Without German $L1$. Horizontal line is base line.

Results for $L1$ classification

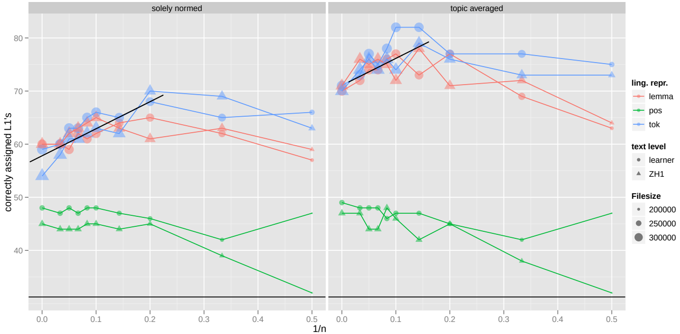


Figure: Classified by $L1$. Without German $L1$. Horizontal line is base line.

observation 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 *L*1 classification

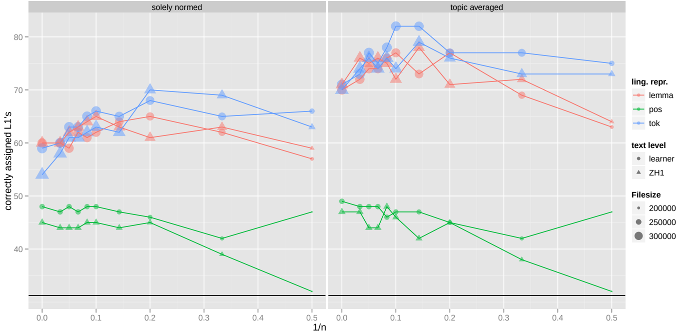


Figure: Classified by *L*1. Without German *L*1. Horizontal line is base line.

Results for *L1* classification

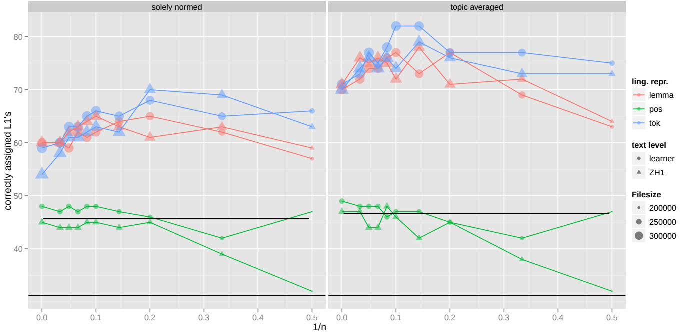


Figure: Classified by *L1*. Without German *L1*. Horizontal line is base line.

observation 2

Filtering out *copied material* does not change much for *POS*

Results for $L1$ classification

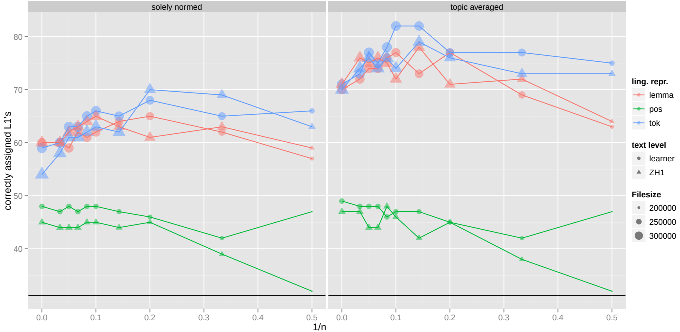


Figure: Classified by $L1$. Without German $L1$. Horizontal line is base line.

Results for $L1$ classification

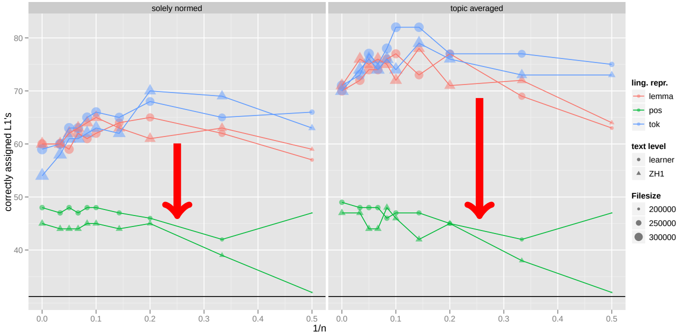


Figure: Classified by $L1$. Without German $L1$. Horizontal line is base line.

observation 3

tok > *lemma* ≫ *POS*

POS is a very reduced text version.

Results for $L1$ classification

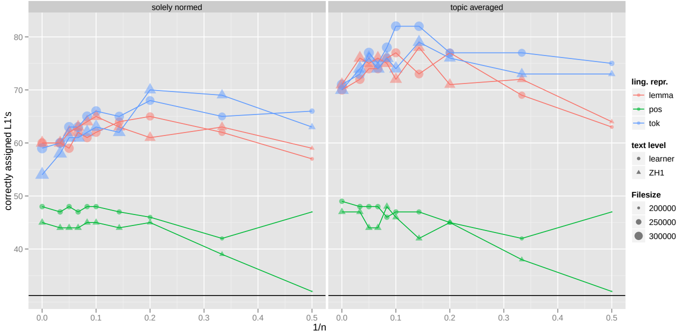


Figure: Classified by $L1$. Without German $L1$. Horizontal line is base line.

Results for *L*1 classification

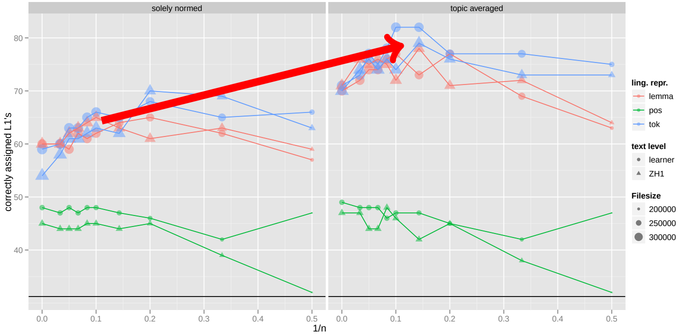


Figure: Classified by *L*1. Without German *L*1. Horizontal line is base line.

observation 4

To average out *topic* helps for *tok* and *lemma*.

$\Rightarrow$ Ignoring *topic* hampers *L*1 classification.

Results for *L*1 classification

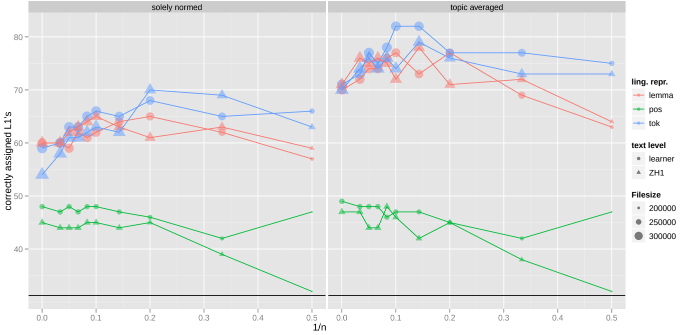


Figure: Classified by *L*1. Without German *L*1. Horizontal line is base line.

Results for $L1$ classification

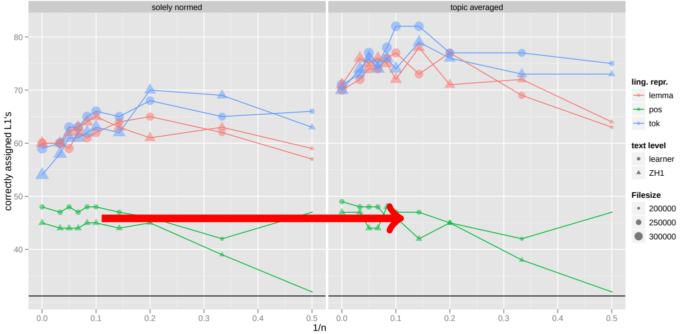


Figure: Classified by $L1$. Without German $L1$. Horizontal line is base line.

observation 5

Again, no such effect for *POS*.

$\Rightarrow$ much less interaction between *topic* and $L1$.

Results for $L1$ classification

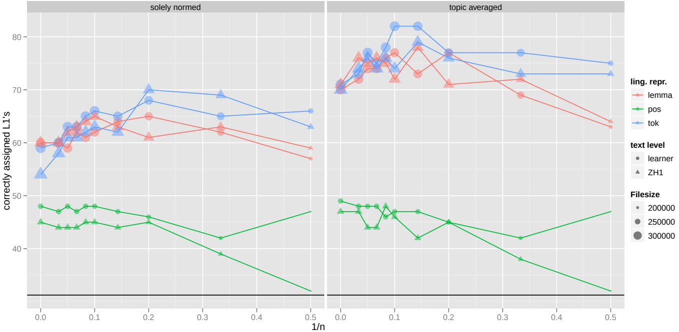


Figure: Classified by $L1$. Without German $L1$. Horizontal line is base line.

Results for *L*1 classification

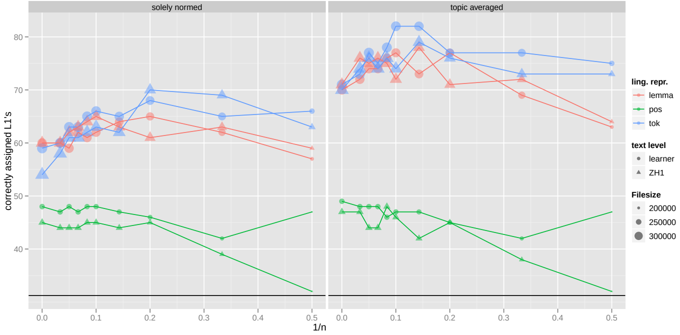


Figure: Classified by *L*1. Without German *L*1. Horizontal line is base line.

observation 6

learner text > ZH1

Correction reduces *L*1 effect. Not so clear for *lemma*.

Results for *topic* classification

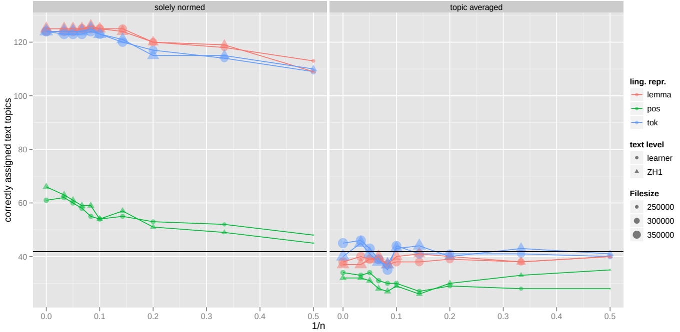
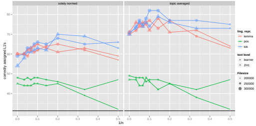


Figure: Classified by text *topic*. With German files. Horizontal line is base line.

Results for *topic* classification



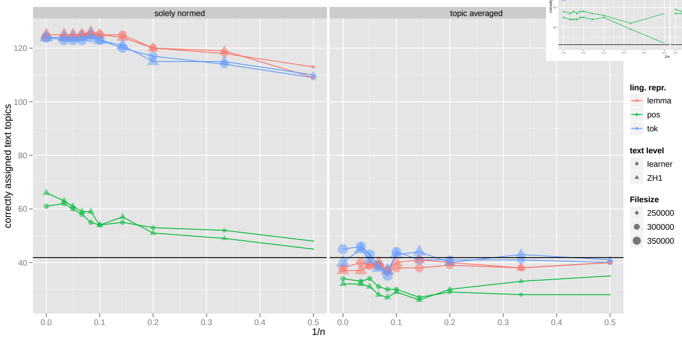
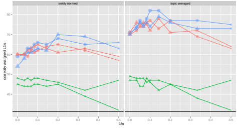


Figure: Classified by text *topic*. With German files. Horizontal line is base line.

Results for *topic* classification



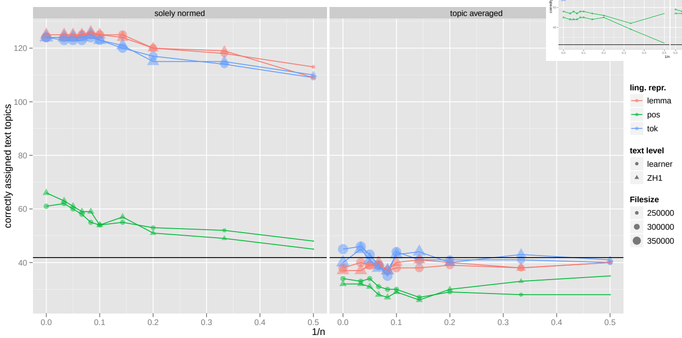


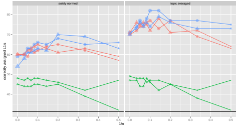
Figure: Classified by text *topic*. With German files. Horizontal line is base line.

observation 1

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

⇒ *Copied material* is not identical with *text topic*.

Results for *topic* classification



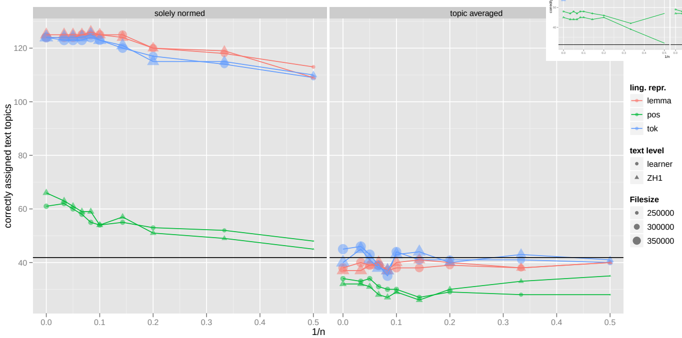
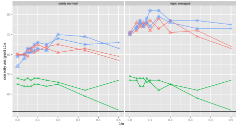


Figure: Classified by text *topic*. With German files. Horizontal line is base line.

Results for *topic* classification



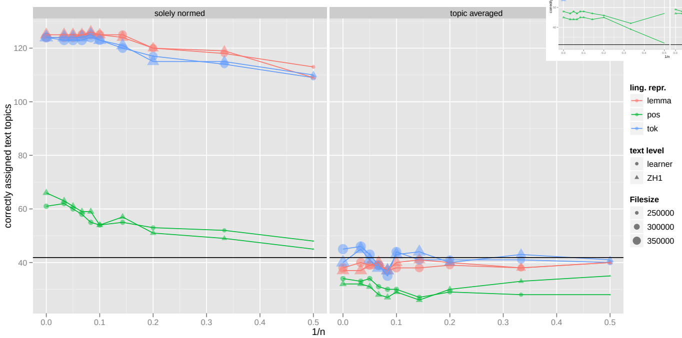
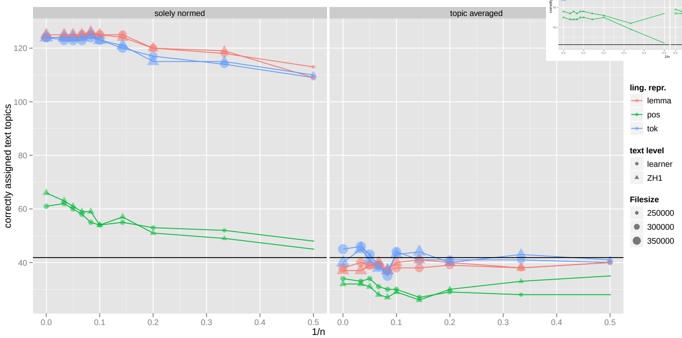


Figure: Classified by text *topic*. With German files. Horizontal line is base line.

observation 2

Again, between $n = 5$ and $n = 10$ is the most interesting strech.

Results for *topic* classification



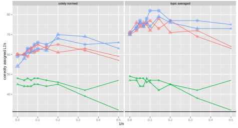
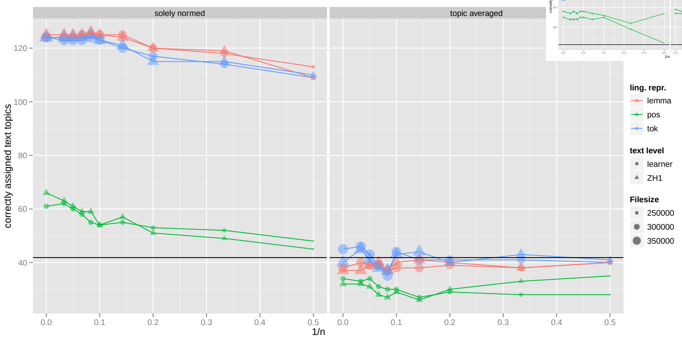


Figure: Classified by text *topic*. With German files. Horizontal line is base line.

Results for *topic* classification



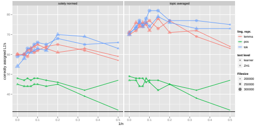
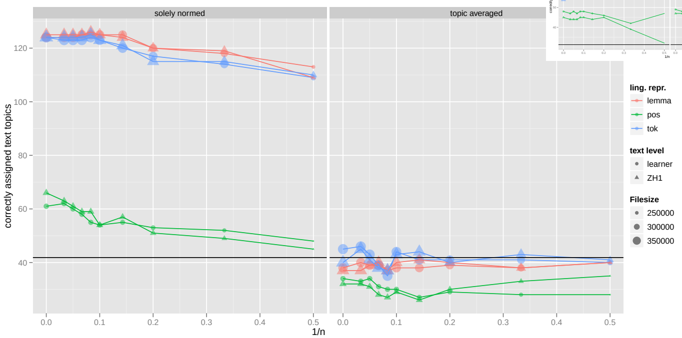


Figure: Classified by text *topic*. With German files. Horizontal line is base line.

observation 3

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

Results for *topic* classification



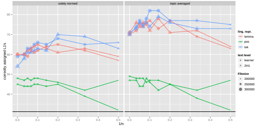
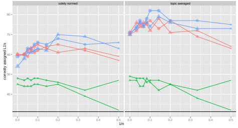


Figure: Classified by text *topic*. With German files. Horizontal line is base line.

observation 3

When more and more is filtered out *POS* nearly hits the base line.
⇒ Much of *L*1 influence in *topic* due to *copied material*. **ALL?**

Results for *topic* classification



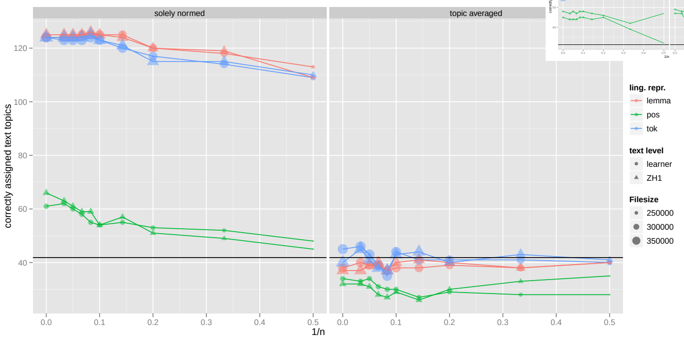


Figure: Classified by text *topic*. With German files. Horizontal line is base line.

Results for *topic* classification

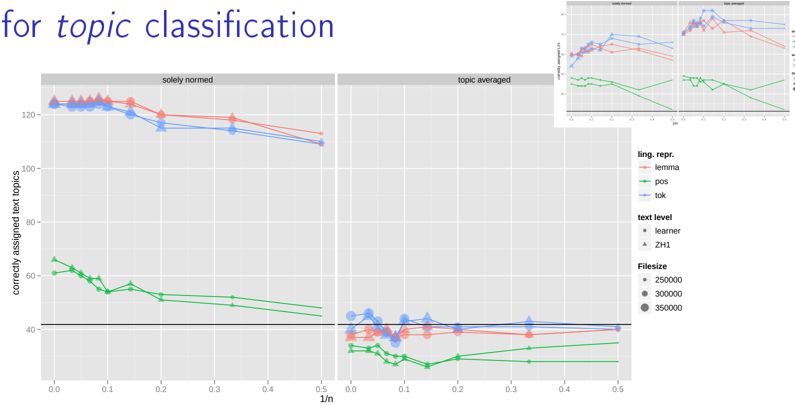


Figure: Classified by text *topic*. With German files. Horizontal line is base line.

observation 4

lemma > *tok* ($\gg$ *POS*)

$\Rightarrow$ *lemma* better for *topic*, *tok* better for *L1* classification.

Results for *topic* classification

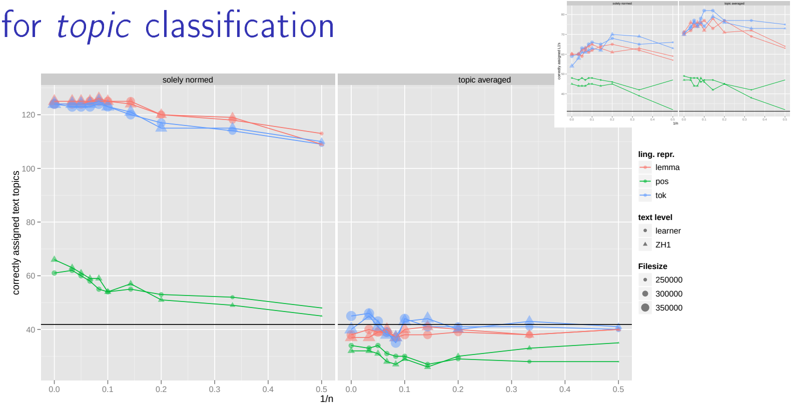
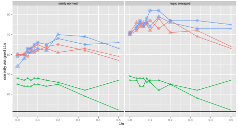


Figure: Classified by text *topic*. With German files. Horizontal line is base line.

Results for *topic* classification



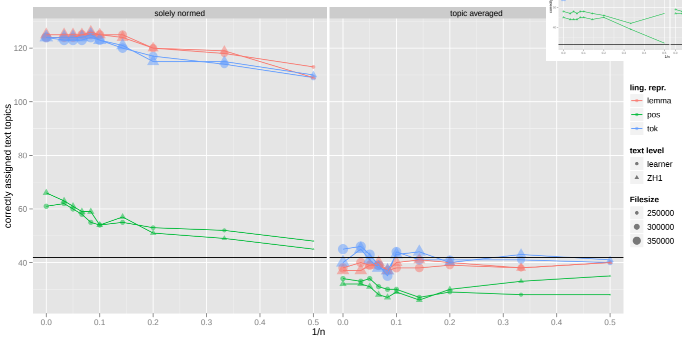


Figure: Classified by text *topic*. With German files. Horizontal line is base line.

observation 5

Our heuristic for *topic* influence reduction works very well:
Performance drops to estimated base line.

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?
  - $\Rightarrow$ 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

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

where

- ϵ 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

$$S = \alpha \cdot \mathit{sameTopic} + \beta \cdot \mathit{sameL}1 + \text{<text specific contributions>} + \epsilon$$

where

- ϵ 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

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

where

  - ϵ 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

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

where

- ϵ 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

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

where

  - ϵ 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

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

  where

  - ϵ is a normally distributed error term.
  - the <text specific contributions> 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:
  - 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

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

  where
  - ϵ is a normally distributed error term.
  - the <text specific contributions> are assumed normally distributed too.
- This (linear mixed) model is fitted.
- The parameters α and β can be compared.

The results

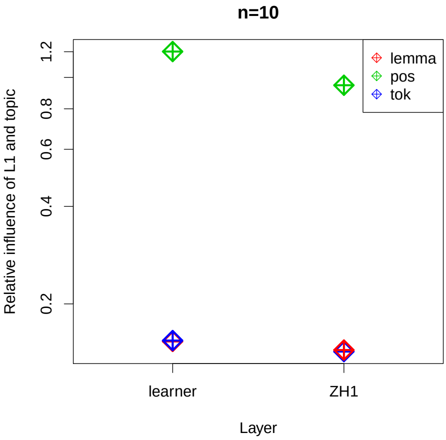


Figure: L1 (β) effect divided by *topic* (α) effect.

observations

1. ***essay topic*** very strong.

The results

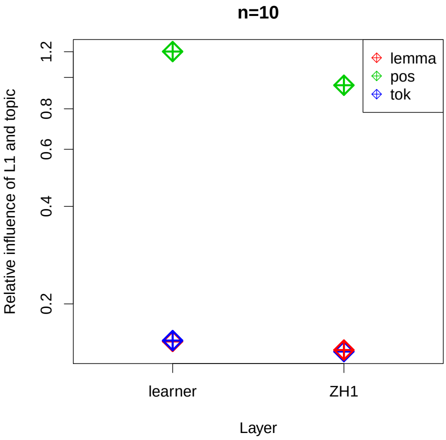


Figure: L1 (β) effect divided by *topic* (α) effect.

observations

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

The results

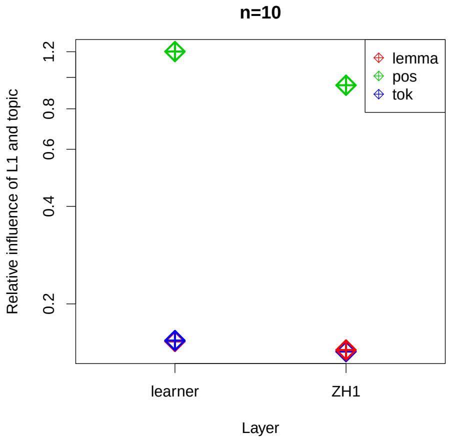


Figure: L1 (β) effect divided by *topic* (α) effect.

observations

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

The results

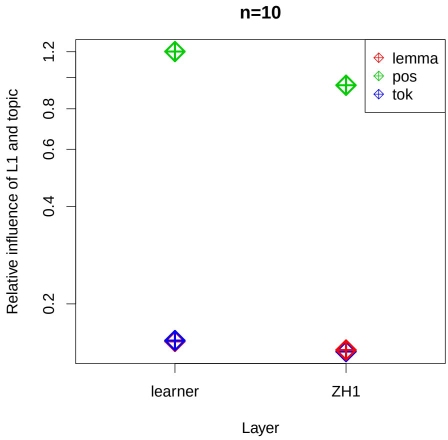


Figure: L1 (β) effect divided by *topic* (α) effect.

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.

The results

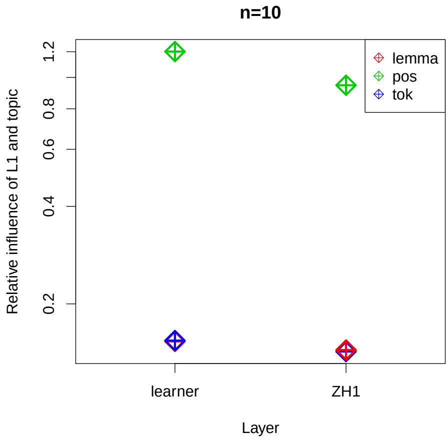


Figure: L1 (β) effect divided by *topic* (α) effect.

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.
5. Removing errors (slightly) weakens L1 influence.

Main results

1. $L1$ has a high and quantifiable influence on our similarity measure S.

Main results

1. $L1$ has a high and quantifiable influence on our similarity measure S.
2. Sensible relations between *learner text* an *Target Hypothesis*.

Main results

1. $L1$ has a high and quantifiable influence on our similarity measure S.
2. Sensible relations between *learner text* an *Target Hypothesis*.
3. Sensible relations between *tok*, *lemma* and *POS*.

Main results

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

What follows in practical terms?

topic

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

copied material

Copied material can distort results.

$\Rightarrow$ 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 I

Cook, Vivian James (2003). *Effects of the second language on the first*. Vol. 3. Second language acquisition. Clevedon: Multilingual Matters. ISBN: 1853596337. URL: http://www.gbv.de/dms/bs/toc/357041879.pdf.

Ellis, Rod, ed. (2009). *The study of second language acquisition*. Oxford applied linguistics. Oxford [u.a.]: Oxford Univ. Press. ISBN: 978 0 19 442257 4.

Golcher, Felix (2007). “A new text statistical measure and its application to stylometry”. In: *Corpus Linguistics 2007*. University of Birmingham.

– (to appear). “Analysing counting suffix trees of natural language texts (preliminary title)”. PhD thesis. Humboldt-Universität zu Berlin.

Granger, Sylviane (2008). “Learner corpora”. In: *Corpus linguistics*. Ed. by Anke Lüdeling et al. Vol. 1. Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science. Berlin, New York: Mouton de Gruyter, pp. 259–275. ISBN: 978-3-11-018043-5.

Literatur II

JojoWong, Sze-Meng et al. (2009). “Contrastive Analysis and Native Language Identification”. In: *Australasian Language Technology Association Workshop 2009*. Ed. by Luiz Augusto Pizzato et al., pp. 53–61.

Juola, Patrick (2004). *Ad-hoc Authorship Attribution Competition*. URL: `http://www.mathcs.duq.edu/~juola/authorship_contest.html` (visited on 03/24/2011).

Koppel, Moshe et al. (2003). “Exploiting Stylistic Idiosyncrasies for Authorship Attribution”. In: *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, pp. 69–72.

Koppel, Moshe et al. (2005). “Automatically Determining an Anonymous Author's Native Language”. In: *Intelligence and Security Informatics*. Lecture Notes in Computer Science. Springer, pp. 209–217. URL: `http://www.springerlink.com/content/rem6vng8r20ebk3q/`.

Lüdeling, Anke et al. (2008). “Das Lernerkorpus Falko”. In: *Deutsch als Fremdsprache* 45.2, pp. 67–73.

Literatur III

Odlin, Terence (2003). “Cross-linguistic Influence”. In: *Handbook on Second Language Acquisition*. Ed. by Catherine Doughty et al. Blackwell, pp. 436–486.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org/`.

Reznicek, Marc et al. (2010). *Das Falko-Handbuch: Korpusaufbau und Annotationen: Version 1.0*. Berlin. URL: `http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko`.

Romaine, Suzanne (2003). “Variation”. In: *The handbook of second language acquisition*. Ed. by Catherine Doughty. Vol. 14. Blackwell handbooks in linguistics. Malden, MA [u.a.]: Blackwell, pp. 409–435. ISBN: 0-631-21754-1.

Literatur IV

Schmid, Helmut (1994). “Probabilistic Part-of-Speech Tagging Using Decision Trees”. In: *Proceedings of the International Conference on New Methods in Language Processing*, pp. 44–49. URL: `http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf`.

Tsur, Oren et al. (2007). “Using Classifier Features for Studying the Effect of Native Language Choice of Written Second Language Words”. In: *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*. ACL, pp. 9–16.

Walter, Maik et al., eds. (2008). *Fortgeschrittene Lernervarietäten: Korpuslinguistik und Zweitspracherwerbsforschung*. Vol. 520. Linguistische Arbeiten. Tübingen: Max Niemeyer Verlag. ISBN: 9783484305205.

Another view on the results

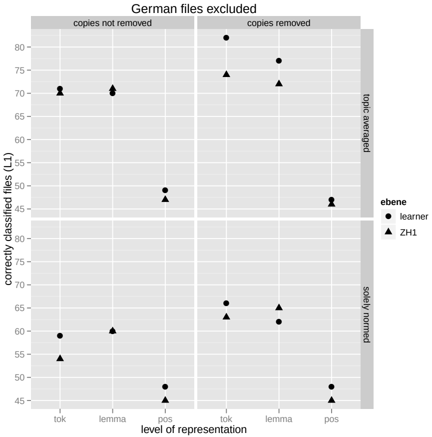


Figure: $L1$ Classification. Maximum at $82/113 = 0.72 \pm 0.09$.

Distribution of right and wrong classifications

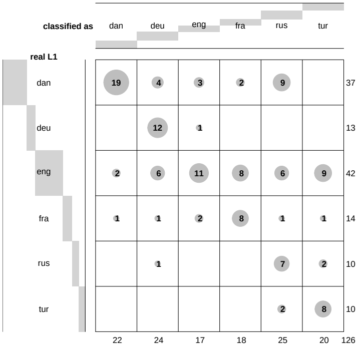


Figure: Raw text.

Distribution of right and wrong classifications

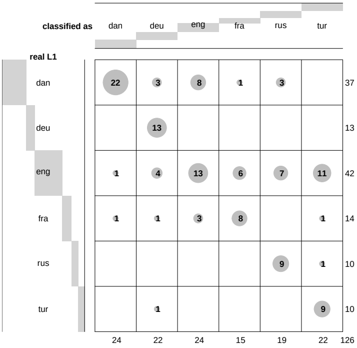


Figure: *title* averaged out.

Distribution of right and wrong classifications

real L1 \ classified as	dan	deu	eng	fra	rus	tur	
dan	22	5	6		4		37
deu		13					13
eng	1	3	21	4	4	9	42
fra		1	1	11		1	14
rus					8	2	10
tur			1			9	10
	23	22	29	15	16	21	126

Figure: copied material removed.

Distribution of right and wrong classifications

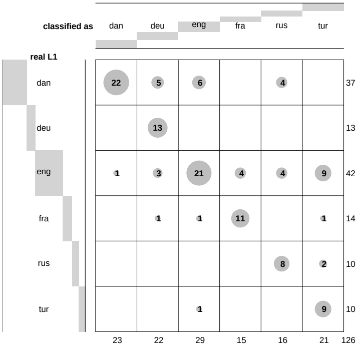


Figure: copied material removed.

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

Distribution of right and wrong classifications

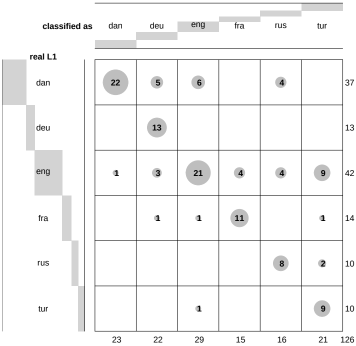


Figure: copied material removed.

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

Distribution of right and wrong classifications

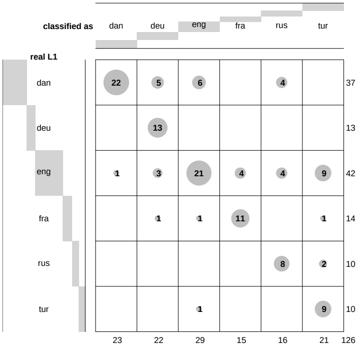


Figure: copied material removed.

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.

Distribution of right and wrong classifications

real L1 \ classified as	dan	deu	eng	fra	rus	tur	
dan	22	5	6		4		37
deu		13					13
eng	1	3	21	4	4	9	42
fra		1	1	11		1	14
rus					8	2	10
tur			1			9	10
	23	22	29	15	16	21	126

Figure: copied material removed.

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)

Distribution of right and wrong classifications

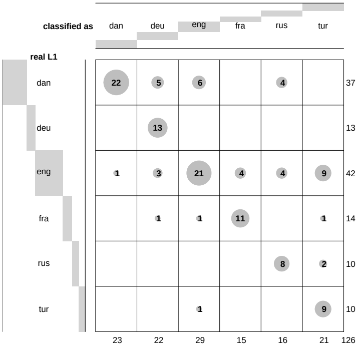


Figure: copied material removed.

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.

Distribution of right and wrong classifications

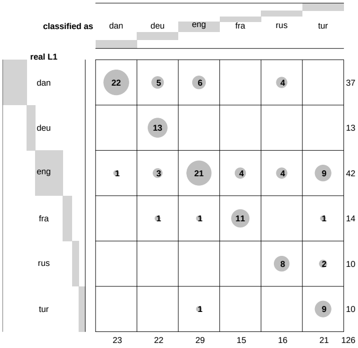


Figure: copied material removed.

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.

Distribution of right and wrong classifications

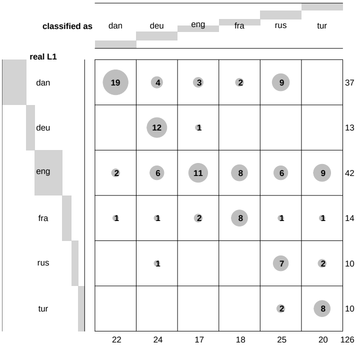


Figure: Raw text.

Distribution of right and wrong classifications

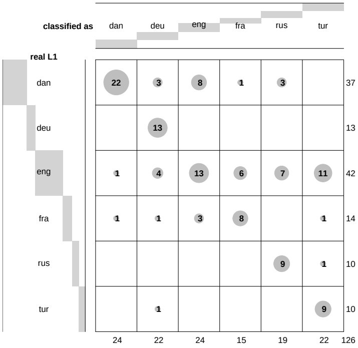


Figure: *title* averaged out.

Distribution of right and wrong classifications

real L1 \ classified as	dan	deu	eng	fra	rus	tur	
dan	22	5	6		4		37
deu		13					13
eng	1	3	21	4	4	9	42
fra		1	1	11		1	14
rus					8	2	10
tur			1			9	10
	23	22	29	15	16	21	126

Figure: copied material removed.

Distribution of right and wrong classifications

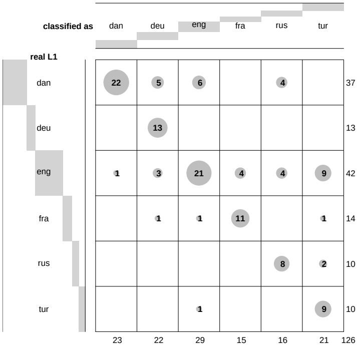


Figure: copied material removed.

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

Distribution of right and wrong classifications

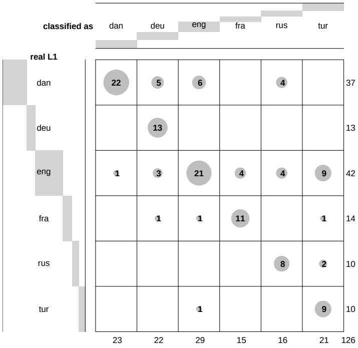


Figure: copied material removed.

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

Distribution of right and wrong classifications

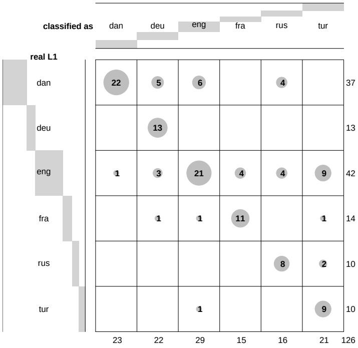


Figure: copied material removed.

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.

Distribution of right and wrong classifications

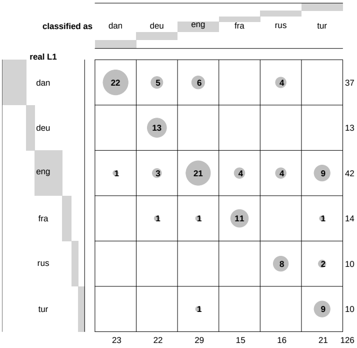


Figure: copied material removed.

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)

Distribution of right and wrong classifications

real L1 \ classified as	dan	deu	eng	fra	rus	tur	
dan	22	5	6		4		37
deu		13					13
eng	1	3	21	4	4	9	42
fra		1	1	11		1	14
rus					8	2	10
tur			1			9	10
	23	22	29	15	16	21	126

Figure: copied material removed.

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.

Distribution of right and wrong classifications

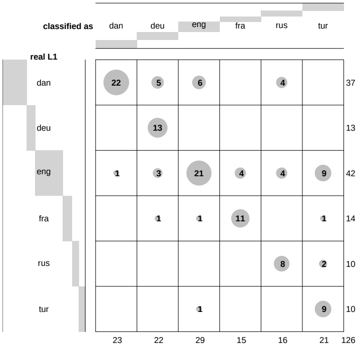


Figure: copied material removed.

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.

11 Norming S

12 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

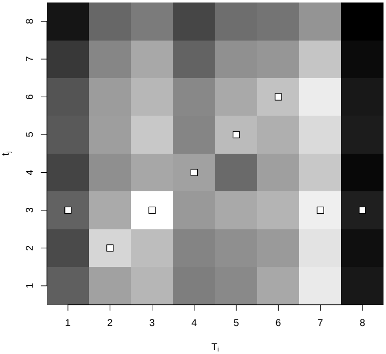


Figure: Dark: low S-values; Light: high S-values.

- Eight Dutch authors[a].
- One training file / one test file.
- Each training file compared with each test file.

=> Training File 8 is the shortest one.

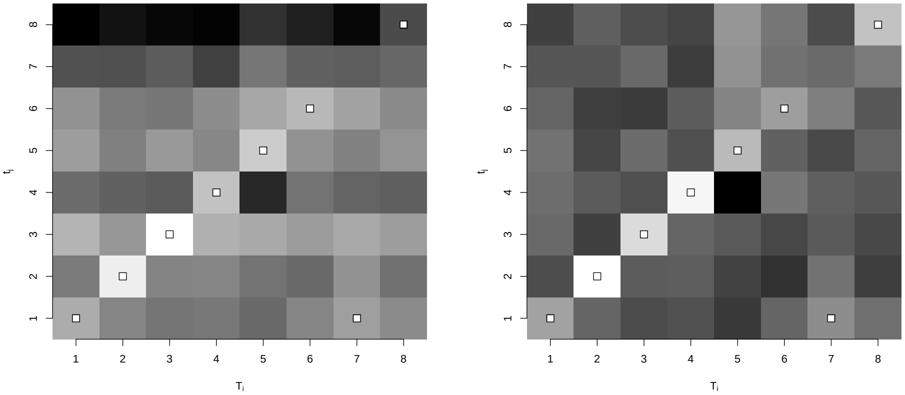
=> Darkest column.

=> lowest S values.

[a] Juola 2004.

Simple: Dividing Columns by their mean.

Averaging out single text dependencies



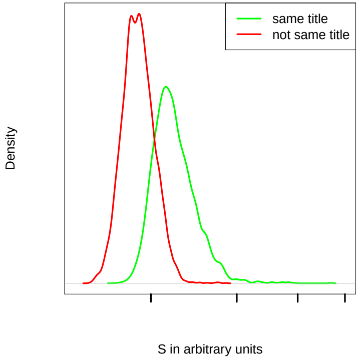
This normed version of S is what we really used.

Distribution of $S(A, B)$ values

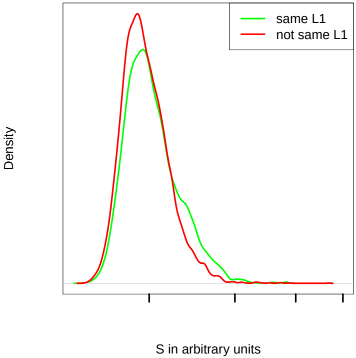
Green: *A* and *B* share *title* or L1

Red: Different *title* or L1.

Same *title* or not?



Same L1 or not?

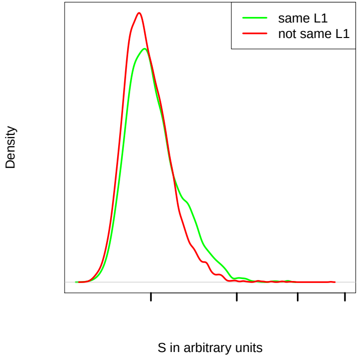


- *title* much stronger than L1.
- But similarity due to L1 is what we are interested in.

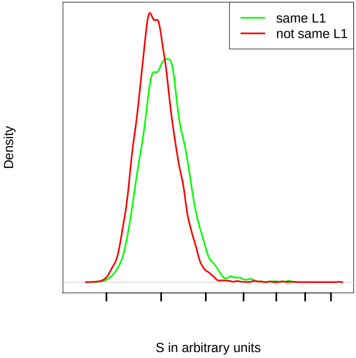
Distribution of $S(A, B)$ values after averaging out *title*

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

with *title*:



title effect removed:

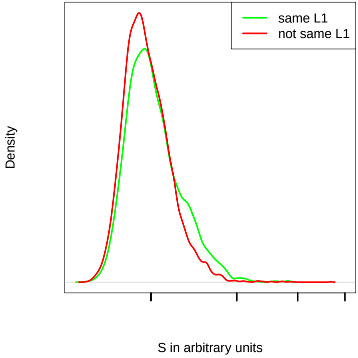


- The difference is much clearer now.
- Classification jumps from 65 to 74 correct decisions (out of 126).

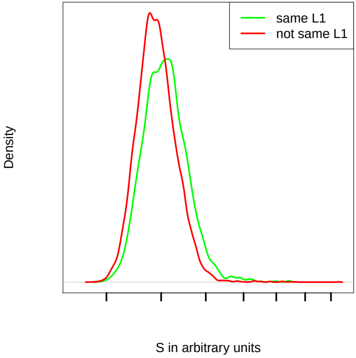
Distribution of $S(A, B)$ values after averaging out *title*

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

with *title:*



title effect removed:

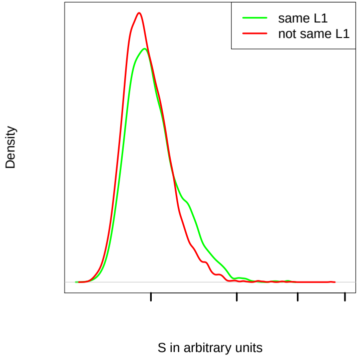


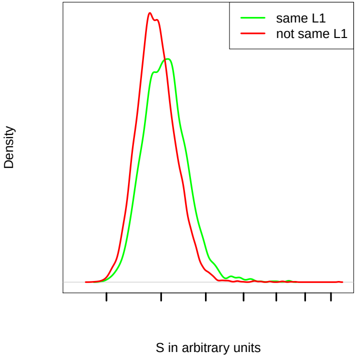
- The difference is much clearer now.
- Classification jumps from 65 to 74 correct decisions (out of 126).
- Suspiciously stretched right tail.

Distribution of $S(A, B)$ values after averaging out *title*

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

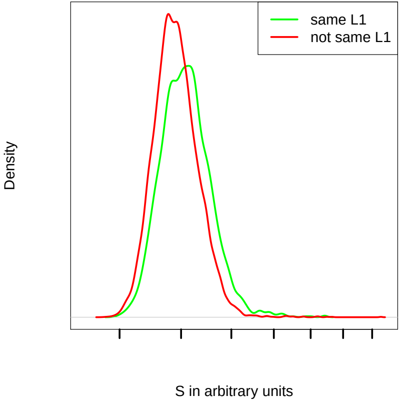
with *title:* *title* effect removed:

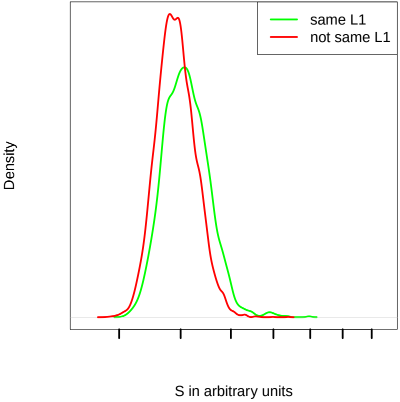




- The difference is much clearer now.
- Classification jumps from 65 to 74 correct decisions (out of 126).
- Suspiciously stretched right tail. ⇒ To this we turn now.

Density plots after removing copied material





- The right tail is greatly reduced.
- Classification results again jump from 74 to 84 correct (from 126).