Stylometry and the interplay of title and L1 in the different annotation layers in the Falko corpus

Felix Golcher & Marc Reznicek

Humboldt-Universität zu Berlin

QITL 4 – 31.03.2011

Coming from second language acquisition research

Learner Corpus Research

1. study of learner language
   - patterns
   - development
   - controlling variables

Coming from second language acquisition research

Learner Corpus Research

1. study of learner language
   - patterns
   - development
   - 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
   - development
   - 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 L1 from her 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 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?

Interlanguage in second language acquisition

Studying language learners as a group, we assume that

Interlanguage in second language acquisition

Studying language learners as a group, we assume that

1. learners of a second/foreign language have a systematic internal grammar (interlanguage: IL)

Interlanguage in second language acquisition

Studying language learners as a group, we assume that

1. learners of a second/foreign language have a systematic internal grammar (interlanguage: IL)
2. IL is different from the native language (L1) grammar

Interlanguage in second language acquisition

Studying language learners as a group, we assume that

1. learners of a second/foreign language have a systematic internal grammar (interlanguage: IL)
2. IL is different from the native language (L1) grammar
3. IL is different from the target language (TL) grammar (Selinker 1972)

Interlanguage in second language acquisition

Studying language learners as a group, we assume that

1. learners of a second/foreign language have a systematic internal grammar (interlanguage: IL)
2. IL is different from the native language (L1) grammar
3. IL is different from the target language (TL) grammar (Selinker 1972)
4. IL has been claimed to be influenced by

Interlanguage in second language acquisition

Studying language learners as a group, we assume that

1. learners of a second/foreign language have a systematic internal grammar (interlanguage: IL)
2. IL is different from the native language (L1) grammar
3. IL is different from the target language (TL) grammar (Selinker 1972)
4. IL has been claimed to be influenced by
   - general learning principles (developmental factors)

Interlanguage in second language acquisition

Studying language learners as a group, we assume that

1. learners of a second/foreign language have a systematic internal grammar (interlanguage: IL)
2. IL is different from the native language (L1) grammar
3. IL is different from the target language (TL) grammar (Selinker 1972)

4. IL has been claimed to be influenced by
   - general learning principles (developmental factors)
   - the structure of the target language

Interlanguage in second language acquisition

Studying language learners as a group, we assume that

1. learners of a second/foreign language have a systematic internal grammar (interlanguage: IL)
2. IL is different from the native language (L1) grammar
3. IL is different from the target language (TL) grammar (Selinker 1972)

4. IL has been claimed to be influenced by
   - general learning principles (developmental factors)
   - the structure of the target language
   - the learner's L1 (transfer)

Interlanguage in second language acquisition

Studying language learners as a group, we assume that

1. learners of a second/foreign language have a systematic internal grammar (interlanguage: IL)
2. IL is different from the native language (L1) grammar
3. IL is different from the target language (TL) grammar (Selinker 1972)

4. IL has been claimed to be influenced by
   - general learning principles (developmental factors)
   - the structure of the target language
   - the learner's L1 (transfer)
   - mode of acquisition, teaching method, learning strategies, psycho-typological aspects, etc.

Transfer as cross-linguistic influence

- Large discussion about what transfer is (Gass et al. 1983; Dechert et al. 1989; Ellis 2009)

- processing mechanism
- learning strategy
- performance/competence phenomenon
- constrains on hypothesis building
- structural borrowing
- etc.

Transfer as cross-linguistic influence

- Large discussion about what transfer is (Gass et al. 1983; Dechert et al. 1989; Ellis 2009)

- processing mechanism
- learning strategy
- performance/competence phenomenon
- constrains on hypothesis building
- structural borrowing
- etc.

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 on different linguistic levels

- Transfer operates on various linguistics levels.

Transfer on different linguistic levels

- Transfer operates on various linguistics levels.
- Many studies have looked at each level independently.

Transfer on different linguistic levels

- Transfer operates on various linguistics levels.
- Many studies have looked at each level independently.
  - phonology (Broselow 1992)
  - morphology (e.g. Dusková 1984; Jarvis 2000)
  - syntax (e.g. Odlin 1990)
  - semantics (e.g. Kellermann 1979)
  - lexicon (e.g. Ringbom 1992)
  - conceptualization (e.g. Stutterheim 1999, Slabakova 2000)
  - etc.

Transfer on different linguistic levels

- Transfer operates on various linguistics levels.
- Many studies have looked at each level independently.
  - phonology (Broselow 1992)
  - morphology (e.g. Dusková 1984; Jarvis 2000)
  - syntax (e.g. Odlin 1990)
  - semantics (e.g. Kellermann 1979)
  - lexicon (e.g. Ringbom 1992)
  - conceptualization (e.g. Stutterheim 1999, Slabakova 2000)
  - etc.

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)

Transfer as overuse/underuse

- Some studies have studied transfer by comparing frequencies of POS-tag n-grams ($n < 5$, see Aarts et al. 1998; Borin et al. 2004)

POS-tag-chains which show a significant overuse/underuse **for a special L1 or subgroup of L1s** indicate transfer effects

Transfer as overuse/underuse

- Some studies have studied transfer by comparing frequencies of POS-tag n-grams ($n < 5$, see Aarts et al. 1998; Borin et al. 2004)

POS-tag-chains which show a significant overuse/underuse **for a special L1 or subgroup of L1s** indicate transfer effects

- In a similar approach Zeldes et al. (2008) study L1 independent IL structures

POS-tag-chains which show a significant overuse/underuse **for all L1s** indicate L2-structural difficulties

Transfer as overuse/underuse

- Some studies have studied transfer by comparing frequencies of POS-tag n-grams ($n < 5$, see Aarts et al. 1998; Borin et al. 2004)

POS-tag-chains which show a significant overuse/underuse **for a special L1 or subgroup of L1s** indicate transfer effects

- In a similar approach Zeldes et al. (2008) study L1 independent IL structures

POS-tag-chains which show a significant overuse/underuse **for all L1s** indicate L2-structural difficulties

- Only short token-based n-grams have been looked at!

Exploiting stylometry to uncover transfer

- A set of studies use machine learning techniques to classify the L1 of the author of IL-texts (ICLE version 1/2)

Transfer effects on classification

If learner text shows special features unique to just one L1-group and distinct from all other L1s, this must be due to transfer (if all other group variables are equally distributed).

- Koppel et al. (2003); Koppel et al. (2005); Tsur et al. (2007); JojoWong et al. (2009); Golcher (to appear)

Exploiting stylometry to uncover transfer

- A set of studies use machine learning techniques to classify the L1 of the author of IL-texts (ICLE version 1/2)

Transfer effects on classification

If learner text shows special features unique to just one L1-group and distinct from all other L1s, this must be due to transfer (if all other group variables are equally distributed).

- Koppel et al. (2003); Koppel et al. (2005); Tsur et al. (2007); JojoWong et al. (2009); Golcher (to appear)
  - L1: Bulgarian, Czech, French, Russian, Spanish
  - L2: English
  - measures based on: errors, function words, rare POS-bi-grams, letter-bi-grams (sub-token)

Exploiting stylometry to uncover transfer

- A set of studies use machine learning techniques to classify the L1 of the author of IL-texts (ICLE version 1/2)

Transfer effects on classification

If learner text shows special features unique to just one L1-group and distinct from all other L1s, this must be due to transfer (if all other group variables are equally distributed).

- Koppel et al. (2003); Koppel et al. (2005); Tsur et al. (2007); JojoWong et al. (2009); Golcher (to appear)
  - L1: Bulgarian, Czech, French, Russian, Spanish
  - L2: English
  - measures based on: errors, function words, rare POS-bi-grams, letter-bi-grams (sub-token)

L1 specific structures in the IL were strong enough to recover the L1 highly above the baseline. ⇒ transfer can be detected by L1-classification.

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

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

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 Corpus

Falko[1] - Error annotated corpus of advanced learners of German (Lüdeling et al. 2008)

[1] http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung-en/falko/standardseite-en

Falko Corpus

Falko[1] - Error annotated corpus of advanced learners of German (Lüdeling et al. 2008)

- written essays & summaries

[1] http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung-en/falko/standardseite-en

Falko Corpus

Falko

Falko[1] - Error annotated corpus of advanced learners of German (Lüdeling et al. 2008)

- written essays & summaries
- L2 essay: 122.789 tokens & L1 essay: 68.485 tokens

[1] http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung-en/falko/standardseite-en

Falko Corpus

Falko[1] - Error annotated corpus of advanced learners of German (Lüdeling et al. 2008)

- written essays & summaries
- L2 essay: 122.789 tokens & L1 essay: 68.485 tokens
- 44 different L1s

[1] http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung-en/falko/standardseite-en

Falko Corpus

Falko

Falko[1] - Error annotated corpus of advanced learners of German (Lüdeling et al. 2008)

- written essays & summaries
- L2 essay: 122.789 tokens & L1 essay: 68.485 tokens
- 44 different L1s
  - largest: Danish, English, Russian, French, Uzbek, Turkish

[1] http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung-en/falko/standardseite-en

Falko Corpus

Falko[1] - Error annotated corpus of advanced learners of German (Lüdeling et al. 2008)

- written essays & summaries
- L2 essay: 122.789 tokens & L1 essay: 68.485 tokens
- 44 different L1s
  - largest: Danish, English, Russian, French, Uzbek, Turkish
- 4 different essay topics (*titles*)

[1] http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung-en/falko/standardseite-en

Falko Corpus

Falko[1] - Error annotated corpus of advanced learners of German (Lüdeling et al. 2008)

- written essays & summaries
- L2 essay: 122.789 tokens & L1 essay: 68.485 tokens
- 44 different L1s
  - largest: Danish, English, Russian, French, Uzbek, Turkish
- 4 different essay topics (*titles*)
  - feminism, wages, criminality, university degree

[1] `http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung-en/falko/standardseite-en`

Falko data subset for classification

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** 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	$A = \texttt{xabay}$	$B = \texttt{bcbabd}$	
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	$A = \mathtt{xabay}$	$B = \mathtt{bcbabd}$	
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	$A = \texttt{xabay}$	$B = \texttt{bcbabd}$	
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	$\log(1 \cdot 0 + 1) = 0$

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	$A = \texttt{xabay}$	$B = \texttt{bcbabd}$	
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	$\log(1 \cdot 0 + 1) = 0$
			$S = \sum = 3.17$

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.

S as an established stylometric measure

Various stylometric tasks have been investigated with S:

- **Translationese**: Have translations their own "style"?
  - Studies in Baroni et al. (2006) have been replicated.
- **Authorship Attribution**: Who wrote the federalist papers? (Golcher 2007)
  - Main stream attribution of disputed essays confirmed.
- **Recoverage of $L1$ in English**:
  - Replication of the mentioned studies Tsur et al. (2007); Koppel et al. (2005) (Golcher 2007; Golcher to appear)

A short reminder of the data.

Falko data subset for classification

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.

Remark

There is no *significant correlation between the* ***essay title*** *and the author's L1.*

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 $\overline{S}_{L1}$.

Proof of concept

Expectation 1

The baseline of random assignments is around $126/6 = 21$.
We expect to be substantially better than this baseline.

Proof of concept

Expectation 1

The baseline of random assignments is around $126/6 = 21$.
We expect to be substantially better than this baseline.

Outcome

With the raw learner texts we get 65 correct assignments out of 126 .
$\Rightarrow$ There is something meaningful going on.

A short reminder of the different representations of Falko.

Falko - 6 representations

Falko

- 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** the grammaticalized version(Reznicek et al. 2010):
       Man denke an die unterschiedlichen Gruppen, die [. . .]

[1] Schmid 1994.

L1 classification

Expectation 2

token representation shows a stronger L1 effect than *lemma*.

Because: *lemma* ignores morphology completely.

L1 classification

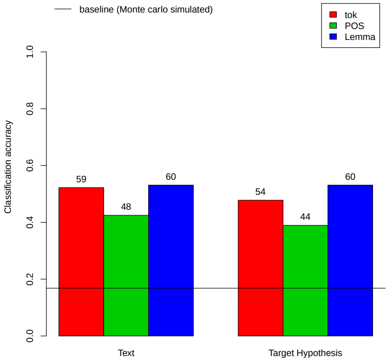


Figure: German L1 texts are disregarded here.

Expectation 2

token representation shows a stronger L1 effect than *lemma*.

Because: *lemma* ignores morphology completely.

L1 classification

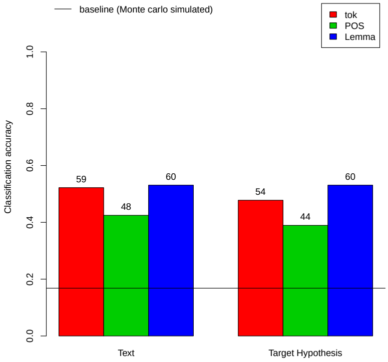


Figure: German L1 texts are disregarded here.

Expectation 2

token representation shows a stronger L1 effect than *lemma*.

Because: *lemma* ignores morphology completely.

Outcome: Big surprise

We could not detect a morphology effect.

L1 classification

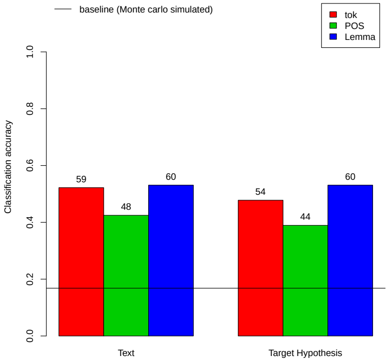


Figure: German L1 texts are disregarded here.

L1 classification

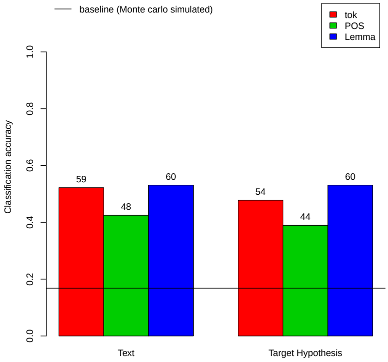


Figure: German L1 texts are disregarded here.

Expectation 3

The grammaticalized ***target hypothesis*** should score somewhat lower than the ***learner*** version.

L1 classification

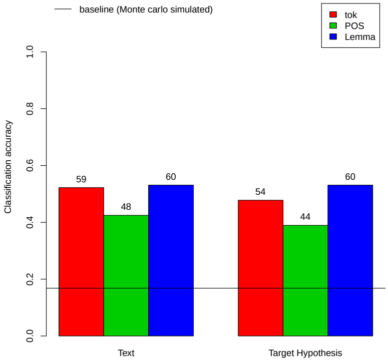


Figure: German L1 texts are disregarded here.

Expectation 3

The grammaticalized ***target hypothesis*** should score somewhat lower than the ***learner*** version.

Outcome

True.
Grammatical error correction lowers accuracy consistently but only minimally.

That's not bad, but didn't we miss some source of similarity?

Another possible influence: Content

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

This is not a newly discovered problem

- This issue is well known from stylometric studies (e.g. Baroni et al. 2006).
- There, usually "content words" are removed. (or similar).
- Rarely any quantitative investigation is tried (but see Clement et al. 2003)

So, how large is this effect? That's what we turn to now...

The same thing with another variable

Now we classify the texts according to **essay title**, instead of **L1**.

classification according to *title*

Expectation 4

We expect a strong **title** effect in the *token* and *lemma* representations.

classification according to *title*

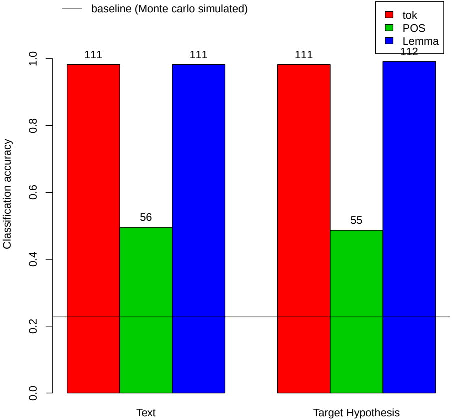
Expectation 4

We expect a strong **title** effect in the *token* and *lemma* representations.

Expectation 5

We do not expect a **title** effect in the *POS* representation.

classification according to *title*



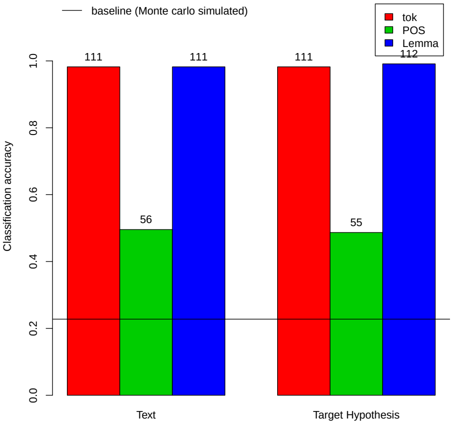
Expectation 4

We expect a strong **title** effect in the *token* and *lemma* representations.

Expectation 5

We do not expect a **title** effect in the *POS* representation.

classification according to *title*



Expectation 4

We expect a strong **title** effect in the *token* and *lemma* representations.

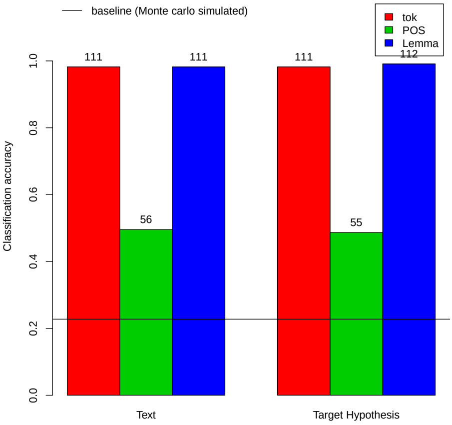
Outcome

nearly perfect.

Expectation 5

We do not expect a **title** effect in the *POS* representation.

classification according to *title*



Expectation 4

We expect a strong **title** effect in the *token* and *lemma* representations.

Outcome

nearly perfect.

Expectation 5

We do not expect a **title** effect in the *POS* representation.

Outcome: Surprise.

Also from the *POS* representation we can recover the **essay title**.

A simple heuristic for filtering out **essay title**

- We divide all $S(A, B)$ in two groups:
  1. A and B have the same *title*.
  2. They have not.
- We compute the mean of each group.
- Each S value is divided by the mean of its group.

Classification results before averaging out *title*

Expectation 6

If we filter out the **essay title**, **L1**-classification improves.

Classification results before averaging out *title*

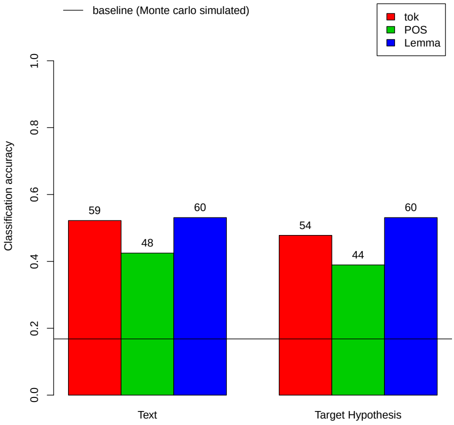


Figure: German L1 texts are disregarded here.

Expectation 6

If we filter out the **essay title**, **L1**-classification improves.

Classification results after averaging out *title*

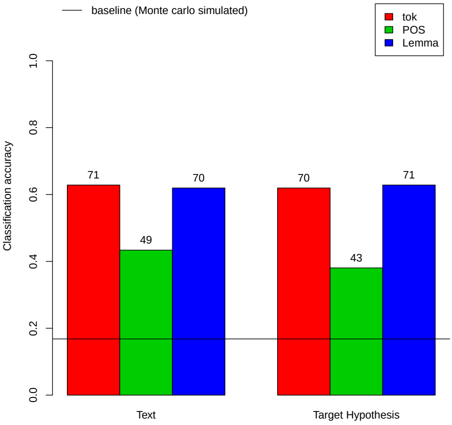


Figure: German L1 texts are disregarded here.

Expectation 6

If we filter out the **essay title**, **L1**-classification improves.

Classification results after averaging out *title*

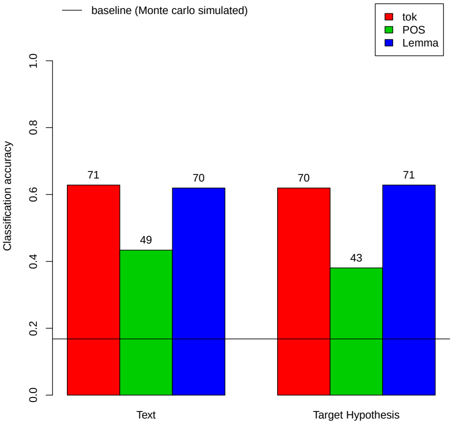


Figure: German L1 texts are disregarded here.

Expectation 6

If we filter out the **essay title**, **L1**-classification improves.

Outcome

True.

Classification results after averaging out *title*

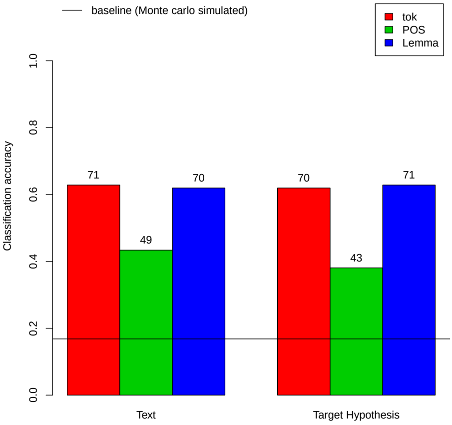


Figure: German L1 texts are disregarded here.

Expectation 6

If we filter out the **essay title**, **L1**-classification improves.

Outcome

True.

Discussion:

Classification results after averaging out *title*

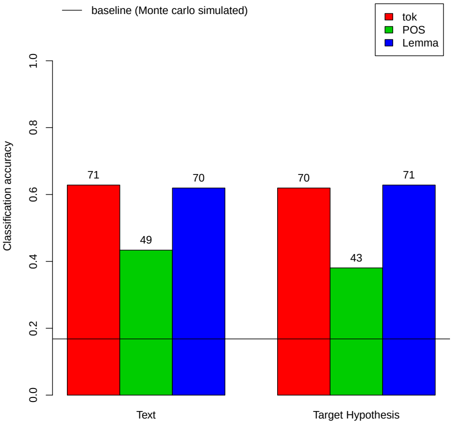


Figure: German L1 texts are disregarded here.

Expectation 6

If we filter out the **essay title**, **L1**-classification improves.

Outcome

True.

Discussion:

1. transfer on lexical choice is much stronger than on syntax. (lemma > POS)

Classification results after averaging out *title*

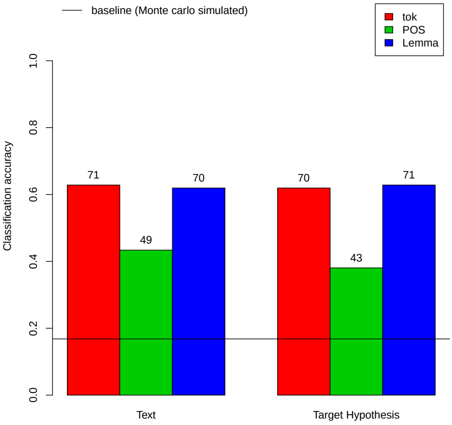


Figure: German L1 texts are disregarded here.

Expectation 6

If we filter out the **essay title**, **L1**-classification improves.

Outcome

True.

Discussion:

1. transfer on lexical choice is much stronger than on syntax. (lemma > POS)
2. We still see no effect of morphology. (lemma = tok)

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 title 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.

Expectation

Expectation 5

If we remove copied material we improve classification performance.

▸ yes, we can remove copied material.

Definition of "copied material"

We use a simple heuristic to identify copied material

Definition (copied material)

A string in text B is copied from text A, if

... it occurs only once in the source text A.

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

Example (set n to 1)

A Do we have beer or do we have wine, Josef?

B Someone must have been telling lies about Josef K.

applying the definition:

" Josef" is copied.

" have b" is not ("have" occurs twice in text A)

Definition of "copied material"

We use a simple heuristic to identify copied material

Definition (copied material)

A string in text B is copied from text A, if

... it occurs only once in the source text A.

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

Example (set n to 1)

A Do we have beer or do we have wine, Josef?

B Someone must have been telling lies about Josef K.

applying the definition:

" Josef" is copied.

" have b" is not ("have" occurs twice in text A)

Definition of "copied material"

We use a simple heuristic to identify copied material

Definition (copied material)

A string in text B is copied from text A, if

... it occurs only once in the source text A.

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

Example (set n to 1)

A Do we have beer or do we have wine, Josef?

B Someone must have been telling lies about Josef K.

applying the definition:

" Josef" is copied.

" have b" is not ("have" occurs twice in text A)

Definition of "copied material"

We use a simple heuristic to identify copied material

Definition (copied material)

A string in text B is copied from text A, if

... it occurs only once in the source text A.

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

Example (set n to 1)

A Do we have beer or do we have wine, Josef?

B Someone must have been telling lies about Josef K.

applying the definition:

" Josef" is copied.

" have b" is not ("have" occurs twice in text A)

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.

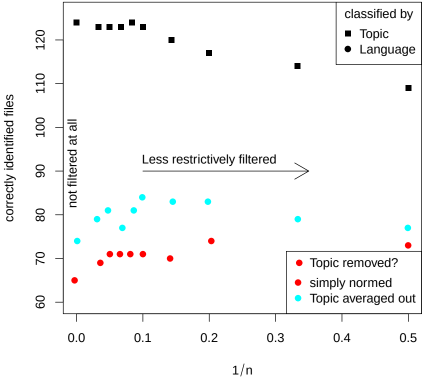
Example

$$n = 10$$

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.

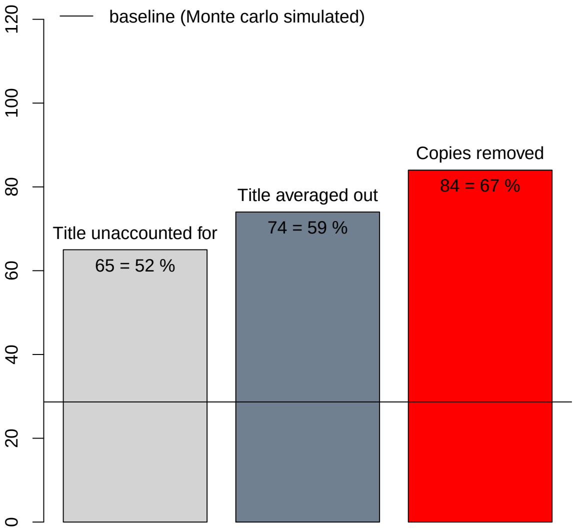
At the end I think, that feminism could help the interests of the women very much, but there are too many people, which harm them concepts of feminism, if they help femininism for wrongs reasons or in wrong situations.

An optimum for the parameter n



- Removing copied material helps identifying L1.
- An approximate optimum is $n = 10$.
- *Title* identification is not hampered.
- Filtering more and more data damps *title* and L1 effect.

Compared classification results



Expectation

Expectation 5

If we remove copied material we improve classification performance.

Outcome

This is indeed the case.

Distribution of right and wrong classifications

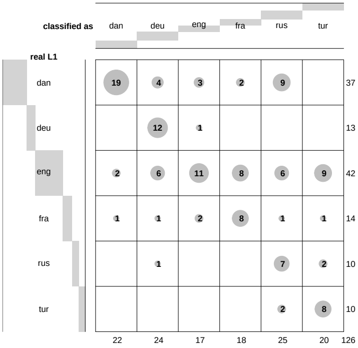


Figure: Raw text.

Distribution of right and wrong classifications

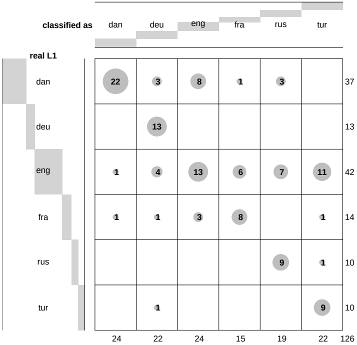


Figure: *title* averaged out.

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

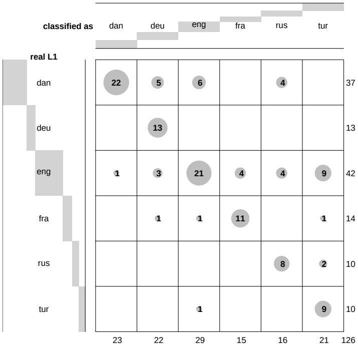


Figure: copied material removed.

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

Distribution of right and wrong classifications

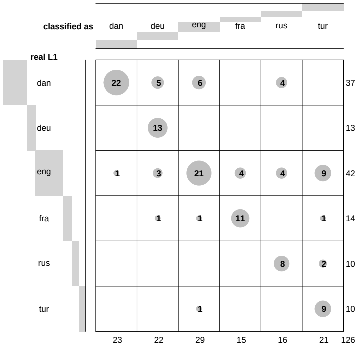


Figure: copied material removed.

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

Distribution of right and wrong classifications

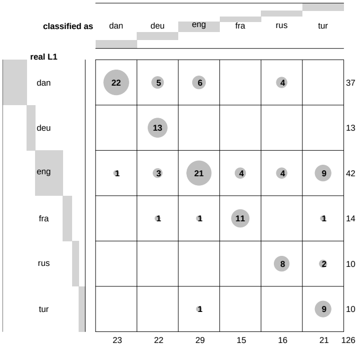


Figure: copied material removed.

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

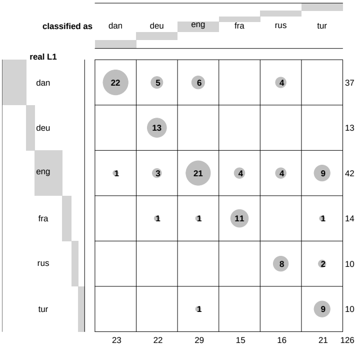


Figure: copied material removed.

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

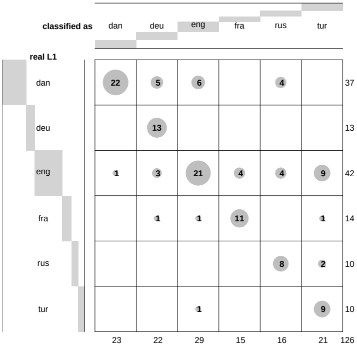


Figure: copied material removed.

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

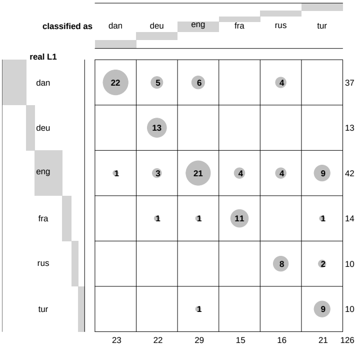


Figure: copied material removed.

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.

Where to go from here?

- Successful classification is a reliable indicator for existing transfer.

  but effect sizes can't be readily quantified.
- The *title* 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 *title* 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:

  sameTitle 1 if A and B share its title, 0 otherwise.

  sameL1 1 if authors of A and B share L1, 0 otherwise.

- Now we set up a model

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

where ϵ is a normally distributed error term.

Building a (linear mixed) model

- For each $S(A, B)$ we construct two variables:

  sameTitle 1 if A and B share its title, 0 otherwise.

  sameL1 1 if authors of A and B share L1, 0 otherwise.

- Now we set up a model

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

where ϵ is a normally distributed error term.

Building a (linear mixed) model

- For each $S(A, B)$ we construct two variables:

  sameTitle 1 if A and B share its title, 0 otherwise.

  sameL1 1 if authors of A and B share L1, 0 otherwise.

- Now we set up a model

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

where ϵ is a normally distributed error term.

Building a (linear mixed) model

- For each $S(A, B)$ we construct two variables:

  sameTitle 1 if A and B share its title, 0 otherwise.

  sameL1 1 if authors of A and B share L1, 0 otherwise.

- Now we set up a model

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

where ϵ is a normally distributed error term.

Building a (linear mixed) model

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

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

  where ϵ is a normally distributed error term.
- This (linear mixed) model is fitted.

Building a (linear mixed) model

- For each $S(A, B)$ we construct two variables:

  sameTitle 1 if A and B share its title, 0 otherwise.

  sameL1 1 if authors of A and B share L1, 0 otherwise.

- Now we set up a model

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

  where ϵ is a normally distributed error term.

- This (linear mixed) model is fitted.
- The parameters α and β can be compared.

The results

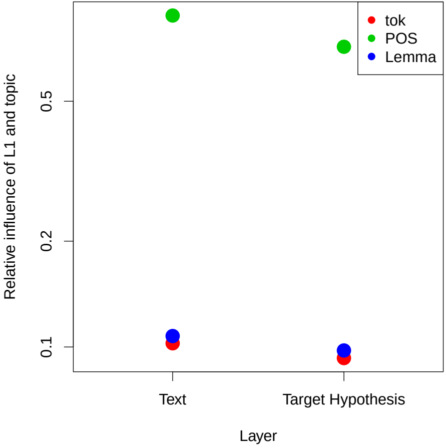


Figure: L1 (β) divided by *title* (α) effect.

observations

1. ***essay title*** always stronger than **L1**: All points below 1.
2. Again, no difference between *token* and *lemma*
3. the **L1** influence in *POS* is much more pronounced.
4. Removing errors (slightly) weakens L1 influence.

What comes out for stylometry

- Stylometric L1 classification is rather successful:
  - Remember how small the data are (66,000 tokens).
  - The method is simple and intuitive.
  - Only substrings, but all substrings are used.
- We can quantify the effects of L1 and *title* or *content*.
- Removal of
  1. *title* influence
  2. copied material

  greatly boosts L1 classification.

That is: S effectively measures L1 induced similarity of learner texts.

What comes out for learner corpus research

1. The presented similarity measure can be used to detect transfer effects.
2. The transfer effect on lexical choice seems considerably stronger than on syntax.
3. Morphological transfer seems to play no significant role in our data.
4. The amount of transfer leading to ungrammaticality seems to be minor.

Warning!

- Learner corpus studies widely ignore the influence of the essay subject (*title*).

  But it's even quite strong on abstract levels such as the Part-of-Speech representation.

Open questions

1. Which substrings in which representation contribute most to the transfer related similarity?

   It is possible to scan the texts character by character and check which contributes what.

Open questions

1. Which substrings in which representation contribute most to the transfer related similarity?

   It is possible to scan the texts character by character and check which contributes what.

2. What is the role of morphology?

   What happens with representation of morphological annotation?

Open questions

1. Which substrings in which representation contribute most to the transfer related similarity?

   It is possible to scan the texts character by character and check which contributes what.

2. What is the role of morphology?

   What happens with representation of morphological annotation?

3. Can we extend the classification to typological similar language groups?

Open questions

1. Which substrings in which representation contribute most to the transfer related similarity?

   It is possible to scan the texts character by character and check which contributes what.

2. What is the role of morphology?

   What happens with representation of morphological annotation?

3. Can we extend the classification to typological similar language groups?
4. Is it possible to use the same method as an indicator of proficiency?

Open questions

1. Which substrings in which representation contribute most to the transfer related similarity?

   It is possible to scan the texts character by character and check which contributes what.

2. What is the role of morphology?

   What happens with representation of morphological annotation?

3. Can we extend the classification to typological similar language groups?
4. Is it possible to use the same method as an indicator of proficiency?
5. How good are the classification results, if all levels are used in combination?

Thank you

Literatur I

Aarts, Jan et al. (1998). “Tag sequences in learner corpora: A key to interlanguage grammar and discourse”. In: *Learner English on computer*. Ed. by Sylviane Granger. Studies in language and linguistics. London [u.a.]: Longman, pp. 132–141. ISBN: 0-582-29883-0.

Baroni, Marco et al. (Sept. 2006). “A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text”. In: *Literary and Linguistic Computing* 21.3, pp. 259–274. ISSN: 0268-1145. DOI: `10.1093/llc/fqi039`.

Borin, Lars et al. (2004). “New wine in old skins? A corpus investigation of L1 syntactic transfer in learner language”. In: *Corpora and language learners*. Ed. by Guy Aston et al. Vol. 17. Studies in corpus linguistics. Amsterdam, Philadelphia: John Benjamins, pp. 67–87. ISBN: 9027222886.

Broselow, Ellen (1992). “Transfer and Universals in Second Language Epenthesis”. In: *Language Transfer in Language Learning*. Ed. by Susan M. Gass et al. John Benjamins Publishing Company, pp. 71–86.

Literatur II

Clement, Ross et al. (2003). "Ngram and Bayesian classification of documents for topic and authorship". In: *Literary and Linguistic Computing* 18.4, pp. 423–447.

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`.

Dechert, Hans Wilhelm et al., eds. (1989). *Transfer in language production*. Norwood, NJ: Ablex Publ. Corp. ISBN: 0893913995.

Dusková, L. (1984). "Similarity-An aid or hindrance in foreign language learning?" In: *Folia Linguistica: Acta Societatis Linguisticae Europaeae* 18, pp. 103–115.

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.

Literatur III

Gass, Susan M. et al., eds. (1983). *Language transfer in language learning*. Issues in second language research. Rowley, Mass.: Newbury House Publ. ISBN: 0883773058.

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

– (to appear). “Repetitions in Text”. PhD thesis. Humboldt-Universität zu Berlin.

Jarvis, Scot H. (2000). “Morphological Type, spatial reference, and language transfer”. In: *Studies in Second Language Acquisition* 22, pp. 535–556. ISSN: 0272-2631.

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).

Literatur IV

Kellermann, Eric (1979). "Transfer and non-transfer: where we are now". In: *Studies in Second Language Acquisition* 2.1, pp. 37–58. ISSN: 0272-2631.

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.

Odlin, Terence (1990). "Word order, metalinguistic awareness, and constraints on foreign language learning". In: *Second Language Acquisition/Foreign Language Learning*. Ed. by Bill VanPatten et al. Channel View Publications Ltd, pp. 95–117. ISBN: 9780585256801.

Literatur V

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 (2010). *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 (visited on 10/12/2010).

Ringbom, Håkan (1992). "On L1 Transfer in L2 Comprehension and production". In: *Language Learning* 42, pp. 85–112. ISSN: 1467-9922.

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 VI

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`.

Selinker, Larry (1972). "Interlanguage". In: *International Review of Applied Linguistics in Language Teaching* 10.3, pp. 209–231. ISSN: 0019042X/2007/045-045.

Slabakova, Roumyana (2000). "L1 transfer revisited: the L2 acquisition of telicity marking in English by Spanish and Bulgarian native speakers". In: *Linguistics* 38.4, pp. 739–770. ISSN: 0024-3949.

Stutterheim, Christiane v. (1999). "How language specific are processes in the conceptualiser?" In: *Represenations and Processes in Language Production*. Ed. by Ralf Klabunde et al. DUV, pp. 153–179.

Literatur VII

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.

Zeldes, Amir et al. (2008). "What's hard? Quantitative evidence for difficult constructions in German learner data". In: *Proceedings of QITL 3*. Helsinki.

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

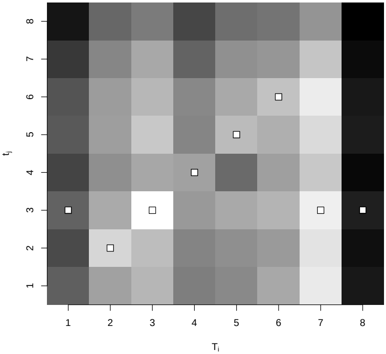


Figure: Dark: low S-values; Light: high S-values.

- 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.

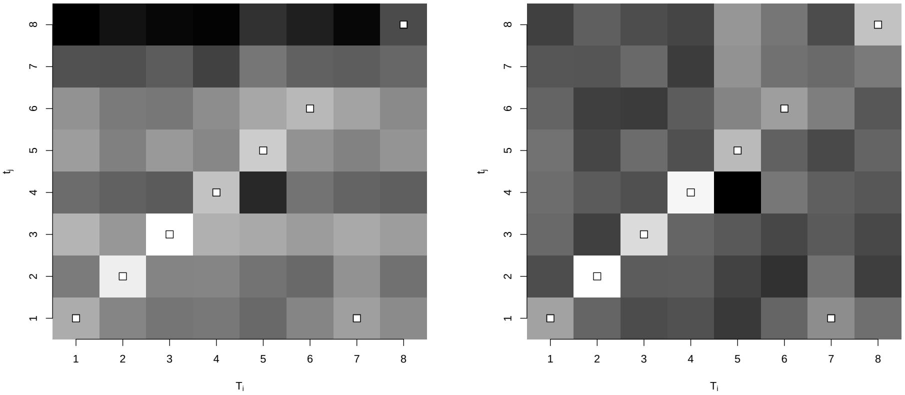
=> Darkest column.

=> 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.

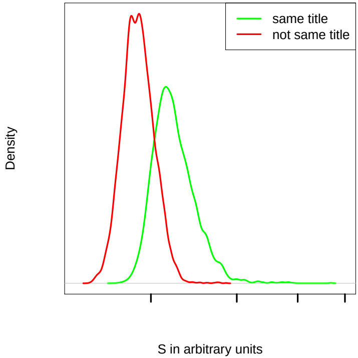
9 Norming S

10 Density plots

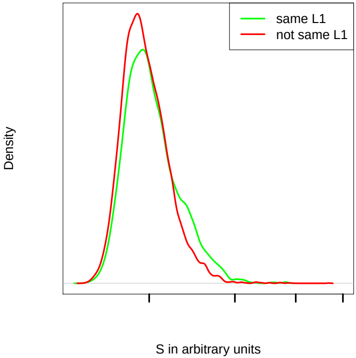
Distribution of $S(A, B)$ values

Green: *A* and *B* share *title* or L1
Red: Different *title* or L1.

Same *title* or not?



Same L1 or not?

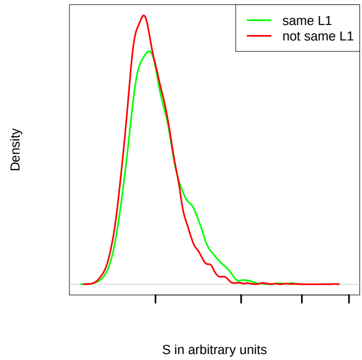


- *title* much stronger than L1.
- But similarity due to L1 is what we are interested in.

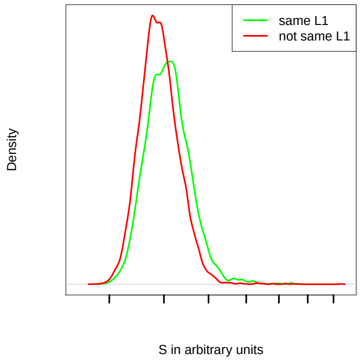
Distribution of $S(A,B)$ values after averaging out *title*

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

with *title:*



title effect removed:

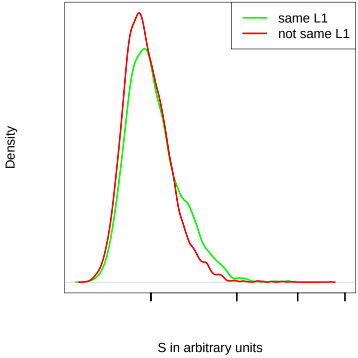


- The difference is much clearer now.
- Classification jumps from 65 to 74 correct decisions (out of 126).

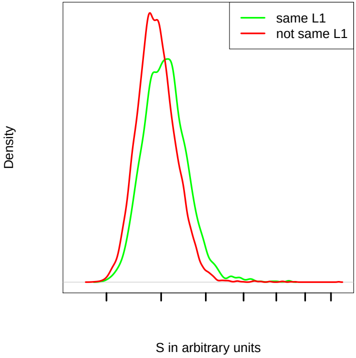
Distribution of $S(A, B)$ values after averaging out *title*

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

with *title:*



title effect removed:

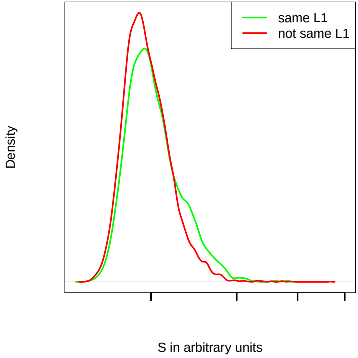


- The difference is much clearer now.
- Classification jumps from 65 to 74 correct decisions (out of 126).
- Suspiciously stretched right tail.

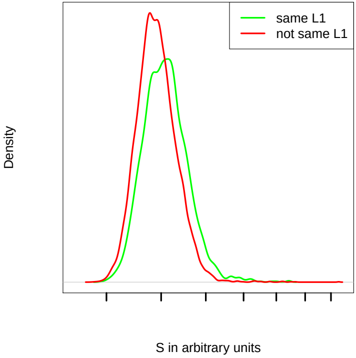
Distribution of $S(A, B)$ values after averaging out *title*

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

with *title*:

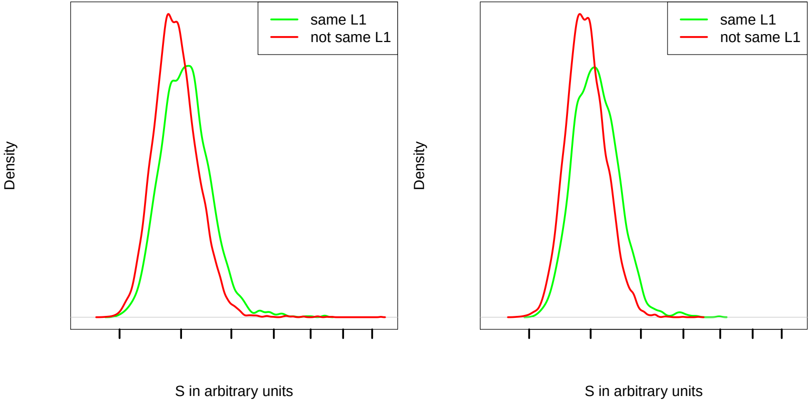


title effect removed:



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