

---

Felix Golcher, Marc Reznicek  
Humboldt Universität zu Berlin  
[felix.golcher@hu-berlin.de](mailto:felix.golcher@hu-berlin.de), [marc.reznicek@staff.hu-berlin.de](mailto:marc.reznicek@staff.hu-berlin.de)

## **Stylometry and the interplay of topic and L1 in the different annotation layers in the FALKO corpus**

### **Introduction**

The process of transfer of structures from the mother tongue (L1) has been shown to take place on different linguistic levels like lexis, syntax, discourse etc. (Ellis 2009:295). Up to now though, it has been difficult to quantify the relative strength of transfer on those different levels at the same time. This will be the main objective of this study. We therefore separately look at the surface level text, sequences of the part-of-speech (POS) tags as a rough representation of the syntactic structure and the sequence of lemmas reflecting mainly the choice of lexical items.

Furthermore we will compare the results not only for the original learner texts, but also for an annotation layer that represents the reconstructed utterances of the learners, thereby enabling us to measure effects that stem from ungrammaticality in the text compared to L1-induced phenomena that do not lead to ungrammatical structures.

We compare argumentative texts on four different topics written by advanced learners of German in the Falko Learner Corpus of German. Our results indicate that besides a known influence of text type and genre (Byrnes et al. 2004) one has to take into account the confounding variable of text topic.

### **Research on n-grams**

Research done on POS tag sequences shows that while certain linguistic structures of learner language display systematic over- and underuse for all mother tongue (L1) groups (Zeldes et al. 2008), many seem to be L1 specific (Aarts et al. 1998, Borin et al. 2004). In those studies the length of the n-grams have been arbitrarily constrained to small numbers (mostly  $n < 5$ ). Our approach takes into account all n-grams found in a text to calculate a similarity between two texts by comparing the two sets of n-grams.

The more n-grams appear in both texts, the higher is the rate of similarity between those two texts. This might allow us to find satisfactory results even in a relatively small corpus like FALKO. The stylometric measure we use is described in Golcher 2007.

## Stylometry

If we train our algorithm on sets of German texts stemming from different mother tongue speakers, we can use this “knowledge” to classify unseen texts into those L1 sets. The more the transparent L1 features show up in the target language text, the better the success of the classification for that language will be. High accuracy would therefore support the claim of very coherent L1 traces which can be interpreted as a result of transfer. Low accuracy would either refute the quality of the method or could be interpreted as a result of L1 independent factors making learner texts too similar to classify them correctly.

If the L1 has an influence on the similarity of two texts, we would expect it to show up at different strengths depending on the amount and type of information included in the text. While the raw text comprises information about lexical choice, morphology, syntax etc., in the chain of automatically annotated part-of-speech tags the lexical information has been lost and only a schematic version of the morpho-syntactic information has been left. In the chain of lemma tags on the other hand morphology has been suppressed while surface linearity and lexical choice have been conserved. Comparing these three layers, one can at least partially separate two kinds of transfer effects: purely lexical effects should not show up on the POS level, while syntactic transfer which cannot be attributed to simple linearity phenomena should be less visible. Stylometric methods have been applied to the task of identifying a writer’s mother tongue before (Koppel et al. 2005a, b; Tsur & Rappoport 2007, Golcher 2007), this however has been done on the surface level only. As an interesting side effect our study partially replicates the results of the papers cited above which are all based on the ICLE corpus (Granger 2002) with data from German English learners.

Since the corpus also includes a target hypothesis (TH) for every learner utterance, which forms a reconstructed target language version of the original utterance (Lüdeling 2008), as well as POS and lemma information for this TH, we can add a comparison of the results for the TH with the original learner texts. Since the TH inherently reflects a mediation between the learner text and a normally produced native speaker text (Reznicek et al. 2010), the L1 effect should be less visible for all three levels of representation. This would then help to evaluate the quality of the methods predictive power of transfer effects.

## The algorithm

The measure used for quantifying the similarity of two texts is based on the frequency distribution of all substrings in both texts. Basically, the products of all those frequency pairs are summed up logarithmically. The immense number of substrings of the texts is handled by storing them as suffix trees. Written as a formula, the similarity measure  $S(T_1, T_2)$  is defined as

$$S(T_1, T_2) = \sum_{\text{all } S} \log[F_{T_1}(s) \cdot F_{T_2}(s) + 1] \quad (1)$$

Here, the sum runs over all substrings  $s$ .  $F_{T1}$  and  $F_{T2}$  denote the frequencies of  $s$  within the two texts. The logarithm ensures that the much higher frequencies especially of very short substrings do not mask the potentially interesting low frequency substrings. Since the logarithm of 1 is 0, all substrings which do not occur in both texts make no contribution to  $S$ , which is as it should be. The addition of one is also needed in order to include substrings occurring once in both texts in  $S$ .

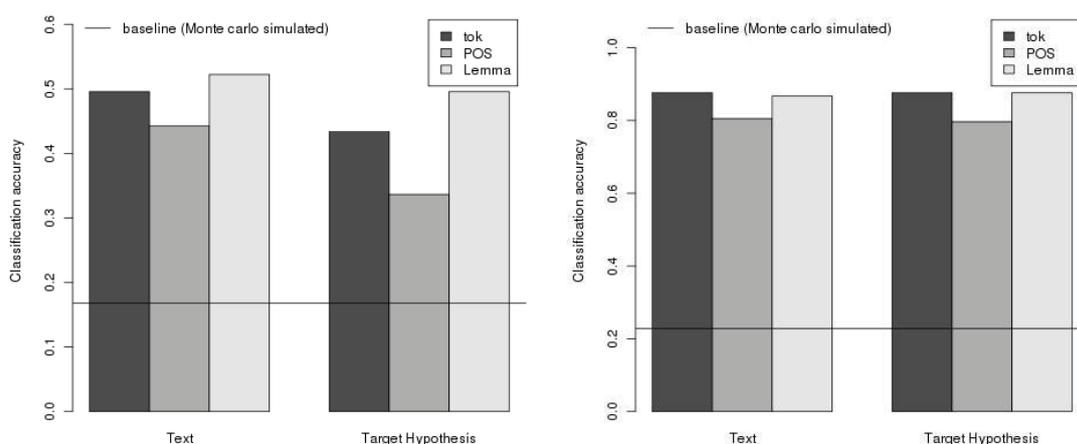
Since the substring frequencies will grow in general as the text gets longer, a normalization procedure is needed. In practice, a simple averaging strategy has turned out to be efficient (see Golcher 2007, where the described method has been shown to be highly competitive within a range of stylometric tasks).

## Empirical analysis

To minimize third language interactions the data was restricted to include texts of non bilingual speakers only. We used essays from the five largest L1 groups only. This left us with 42 English, 37 Danish, 14 French, 10 Russian, and 10 Turkish native speakers. All files together consisted of just below 60,000 tokens.

For all possible text pairs,  $S$  was computed. This yields a symmetric 126 by 126 matrix. As hinted above, this matrix is normalized by dividing each cell by the mean of the respective row and column (Golcher 2007). This removes all text specific contributions to  $S$ , most notably the impact of text length. The variance due to the topic of the essays is still rather large. Thus we split all  $S$  values into two sets, where both contributing texts do or do not belong to the same topic. The  $S$  values of both sets were divided by their means. Then, each file in turn was classified into the L1 group with which it shared the highest mean  $S$  values, doubly normalized as described. In a second run, the texts were classified according to their topic in the same way. Here, we used the simple normalized  $S$  values without regard to the L1 influence. This is justified by the relatively small L1 effect compared to topic. Figure 1 shows the results.

FIGURE 1. CLASSIFICATION ACCURACIES FOR L1 (LEFT PANEL) AND TOPIC (RIGHT PANEL). FOR CLASSIFYING L1, THE TOPIC INFLUENCE WAS MITIGATED BY THE HEURISTICS DESCRIBED IN THE TEXT.



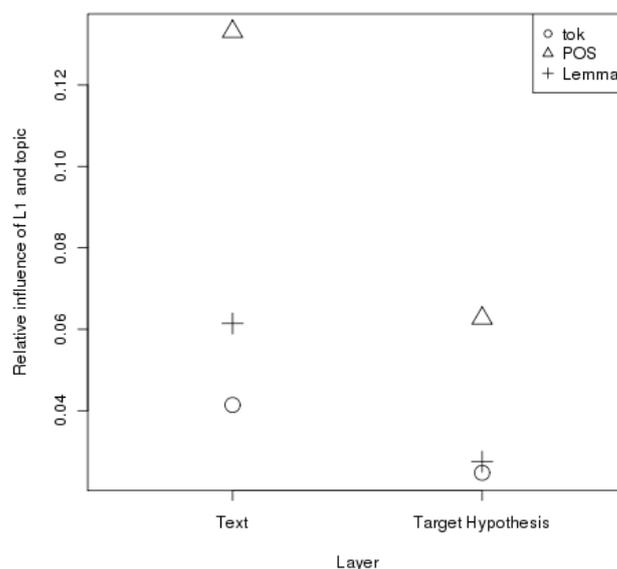
From the left panel of Figure 1 we can draw the following conclusions about the predictability of L1:

1. The overall results of correct predictions for the L1 is significantly higher than random, with a  $p$  value close to 0.
2. There is a difference in the performance for the different levels of representation: Lemma is best, followed by tokens and POS.
3. The same pattern is repeated for the target hypothesis, but at a lower level.

The accuracy for topic classification on the other hand is very high. Predictably, it is lowest in the POS representation, even if it is still astonishingly high. This striking persistence of the topic predictability can be seen as a hint to a spurious effect: The topical influence could be transferred from the lexical level, for example by identical material in different essays, copied from the essay title. This spuriousness cannot be expected to apply to the more interesting L1 influence in any way.

From these results we see how the classification accuracy of L1 and topic varies with regard to the two variables of text/TH and linguistic representation. The baselines for L1 and topic classifications differ though, and slightly different forms of  $S$  were used in both cases. This makes it hard to directly compare the impact of L1 and topic on  $S$ , that is on text similarity. Thus we fitted a linear mixed model (Pinheiro & Bates 2000) to the data. Assuming a linear dependency of  $S$  on the other variables is an approximation of course, but such a model allows us to quantify L1 and topic effects at the same time. The model has to be a mixed one in order to account for the crossed random effects of the individual essays. The results are shown in Figure 2.

FIGURE 2. COMPARING EFFECT SIZES AS GIVEN BY THE FIT TO A LINEAR MIXED MODEL. DISPLAYED IS THE ESTIMATE FOR THE L1 EFFECT DIVIDED BY THE ESTIMATE FOR THE TOPIC EFFECT.



We can summarize them as follows:

1. The effect of topic is always much more pronounced than the effect of L1. Nevertheless, sameness of L1 was always a significant factor within the model.

2. In the POS representation, L1 is relatively much more influential. Thus, even if a topic effect is still rather strong in this representation, the L1 effect resides much more in this structural representation of the text.
3. The same pattern is repeated in the target hypothesis, but the L1 effect is much weaker in comparison to the topic effect. This is in line with expectations: The target hypothesis aims at correcting errors and errors can be assumed to be L1 specific. And, since the lexical content is supposed to change rather little between the two corpus layers, the topic effect should be very similar.

## Conclusion

First, the very high accuracy of the L1 classification strongly supports the claim that L1 dependent transfer effects are present in the learner texts.

The separation of the representation levels shows that transfer dominates on the lexical level. Still there are clear indications of a comparable L1 transfer on the syntactic level. Since the TH corrects mainly morphosyntactic errors, this shows that there are at least in part L1 specific errors that contribute to the coherence of the L1 groups. On the other hand the numbers make it clear that transfer is a lot stronger on levels that do not lead to ungrammatical structures in the strict sense.

Interestingly, the L1 effect on the lemma level is even higher than on the token level, which means that by subtracting morphological clues the L1 effect grows. This can be explained considering that morphology can be related to cues like tense and modality. In fact, topics seem to differ in their likelihood to trigger certain tenses and modalities therefore diminishing L1 influence.

The numbers in the last figure allow the conclusion that the influence of text topic on forms of expression is much larger than has been considered so far, even with regard to syntactic structures. Future studies should therefore control the data not only for text type, register and genre but for topic as well.

## References

- Aarts, Jan and Granger, Sylviane 1998. Tag sequences in learner corpora. a key to interlanguage grammar and discourse. In *Learner English on computer*. [Studies in language and linguistics], Granger, Sylviane (ed.), 132-141. London: Longman.
- Borin, Lars and Prütz, Klas 2004. New wine in old skins? A corpus investigation of L1 syntactic transfer in learner language. In *Corpora and language learners*. [Studies in corpus linguistics 17], Aston, Guy, Bernardini, Silvia and Stewart, Dominic (eds.), 67-87. Amsterdam, Philadelphia: John Benjamins.
- Byrnes, Heidi and Sprang, Katherine A. 2004. Fostering advanced L2 literacy. A genre-based, cognitive approach. In *Advanced foreign language learning. A challenge to college programs*. [Issues in language program direction 2003], Byrnes, Heidi, Hiram, Maxim H. (eds.), Boston MA: Thomson/Heinle.
- Ellis, Rod 2009. *The study of second language acquisition*. 2. ed. Oxford: Oxford Univ. Press.
- Falko: <http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung-en/falko/standardseite-en>.
- Golcher, Felix 2007. A new text statistical measure and its application to stylometry. *Proceedings of Corpus Linguistics*. University of Birmingham. 2007. URL: <http://amor.cms.hu-berlin.de/~golcherf/cl07.pdf>.
- Granger, Sylviane, Dagneaux, Estelle and Meunier, Fanny 2002. *The International Corpus of Learner English. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.
- Granger, Sylviane 2008. Learner corpora. In *Corpus linguistics. An international Handbook* [Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 29 1], Lüdeling, Anke and Kytö, Merja (eds.), 259-275. Berlin, New York: Mouton de Gruyter.
- Koppel, Moshe, Schler, Jonathan, Zigdon, Kfir 2005a. Automatically Determining an Anonymous Author's Native Language. *Intelligence and Security Informatics*. Lecture Notes in Computer Science. Springer. 209-217.

- Koppel, Moshe, Schler, Jonathan, Zigdon, Kfir 2005b. Determining an Author's Native Language by Mining a Text for Errors. In *Proceedings of KDD '05*. Chicago IL.
- Lüdeling, Anke, Doolittle, Seanna, Hirschmann, Hagen, Schmidt, Karin and Walter, Maik. Das Lernerkorpus Falko 2008. *Deutsch als Fremdsprache* 45(2), 67-73.
- Pinheiro, Jose, Bates, Douglas 2000. *Mixed Effects Models in S and S-Plus*. Springer, Berlin.
- Reznicek, Marc, Walter, Maik, Schmidt, Karin, Lüdeling, Anke, Hirschmann, Hagen and Krummes, Cedric 2010. *Das Falko-Handbuch. Korpusaufbau und Annotationen*. Version 1.0. Berlin: Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin. URL: <http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko> [15 October 2010].
- Tsur, Oren, Rappoport, Ari 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. 9-16.
- Zeldes, Amir, Lüdeling, Anke, Hirschmann, Hagen 2008. What's Hard? Quantitative Evidence for Difficult Constructions in German Learner Data. In *Proceedings of QITL3*, Arppe, Antti, Sinnemäki, Kaius and Nikanne, Urpo (eds.), 74-77. Helsinki. URL: [http://www.ling.helsinki.fi/sky/tapahtumat/qitl/Abstracts/Zeldes\\_et\\_al.pdf](http://www.ling.helsinki.fi/sky/tapahtumat/qitl/Abstracts/Zeldes_et_al.pdf).
-