Native Language Identification Across Text Types: How Special Are Scientists?

Sabrina Stehwien*
Universität Stuttgart, Germany

Sebastian Padó*
Universität Stuttgart, Germany

Native Language Identification (NLI) is the task of recognizing the native language of an author from text that they wrote in another language. In this paper, we investigate the generalizability of NLI models among learner corpora, and from learner corpora to a new text type, namely scientific articles. Our main results are: (a) the science corpus is not harder to model than some learner corpora; (b) it cannot profit as much as learner corpora from corpus combination via domain adaptation; (c) this pattern can be explained in terms of the respective models focusing on language transfer and topic indicators to different extents.

1. Introduction

Native Language Identification (NLI) is the task of recognizing an author's native language (L1) from text written in a second language (L2). NLI has found substantial interest in computational linguistics over the last years. It is of interest in various contexts, including language learning, where native speakers tend to commit specific errors (Odlin 1989), data collection in corpus linguistics (McEnery and Baker 2003), the detection of phishing attacks (Estival et al. 2007) or authorship analysis in forensics (Perkins 2015).

Consequently, numerous models for the NLI tasks have been proposed. Almost all models couch NLI as a classification task, where the classes are the potential native languages (L1s) of the author and the features are supposed to model the effects of the author's L1 on the L2. A wide range of feature sets with various degrees of linguistic sophistication has been proposed, many inspired by the related task of authorship identification. They range from function words and structural features (Tetreault et al. 2012; Wong and Dras 2011; Bykh and Meurers 2014; Cimino et al. 2013) to n-grams over characters, words and POS tags (Tsur and Rappoport 2007; Brooke and Hirst 2011; Bykh and Meurers 2012).

One major problem that NLI shares with many other NLP tasks is the *dependence on text type (domain and genre)* of any learned models. For instance, the widely used International Corpus of Learner English (ICLE, Granger et al. 2009) consists of semi-formal learner essays on personal experiences, which are presumably of limited use to induce an NLI model useful on, say, phishing emails. This problem is arguably compounded by concerns about the *status of the NLI models' features*. Ideally, one would like the features to be interpretable in terms of *language transfer* or *language interference*, that is, they should encode influences that the authors' L1 has on their L2 writing and/or translating (Baker 1993; Truffaut 1997; Cardinaletti and Garzone 2005). Recently, Malmasi and Dras (2015) addressed this issue in a large cross-corpus evaluation study, and present

* Institut für Maschinelle Sprachverarbeitung, Pfaffenwaldring 5b, Universität Stuttgart, 70569 Stuttgart, Germany. E-mail: {stehwien,pado}@ims.uni-stuttgart.de

an analysis of frequently used words and syntactic patterns observed in texts by Japanese L1 speakers.

Regarding the ICLE, Brooke and Hirst (2011) claim that the corpus suffers from an inherent *topic bias* due to its construction from essays written in a small set of language courses. As a consequence, the author's L1 correlates strongly with the topics of their essays. If a model therefore learns to distinguish topics rather than L1, there is grounds for concern that the model will also fail to generalize well; compare also, e.g., Petrenz and Webber 2011. (Bykh and Meurers (2012) report, however, that their ICLE models generalize well to other learner corpora.)

In this article, we investigate the issue of text type dependence and feature interpretation on a novel domain–genre combination which, to our knowledge, has so far not been considered in the NLI literature, namely *scientific articles*. Scientific texts do not only follow fairly fixed structures, they also exhibit specific constraints at the level of vocabulary: lexical choice is limited by the use of domain terminology to refer to relevant concepts, and very deliberate choices regarding semantic aspects like modality, discourse markers, or hedges in order to to express statements precisely (Teufel and Moens 2002; Swales 2004; Hyland 2009). Language interference effects are also well documented in academic text, despite the fact that such texts are sometimes proofread by native speakers (Galvão 2009; Olohan and Salama-Carr 2011). Thus, scientific texts appear to be a very interesting and challenging text type for Native Language Identification.

Our article makes three contributions. First, we describe the construction of an L1-labeled corpus of English scientific papers from our own field, computational linguistics, called ACL-NLI. Second, we investigate the usefulness of *domain adaptation techniques* to improve the performance of NLI models on the ACL-NLI and traditional, "generic" NLI corpora. Third, we perform an analysis of the features learned by these models. We find that quantitatively, within-text type models for NLI do as well on the ACL as on more traditional NLI corpora. However, across-genre domain adaptation methods have a harder time generalizing to the ACL-NLI than to generic NLI corpora. The reason is revealed by a qualitative feature analysis: cross-text type learning can remove topic bias on generic NLI corpora, but not on the ACL-NLI, where many features are related to the preferred research topics of different countries.

2. Datasets for Native Language Identification

In our study, we use subsets of three existing learner corpora, plus one new scientific corpus whose construction is described in more detail below (Table 1). We include the seven languages that are in the intersection of all datasets (German/DE, Spanish/ES, French/FR, Italian/IT, Japanese/JP, Turkish/TR, Chinese/ZH). To obtain an NLI task

Table 1
Subsets of the original NLI corpora used in this article

Corpus	# Docs/L1	Avg # Tokens/Doc	Characterization
ICLE	251	612	Learner (specific, well-controlled)
TOEFL11	1100	348	Learner (varied, but still controlled)
Lang-8	176	731	Learner (free, social media-like)
ACL-NLI	54	3850	Scientific text

Table 2
Statistics on the relevant native languages in the original ICLE, Lang-8 and ACL corpora

L1	ICLE	Lang-8	ACL-NLI
DE	437	706	691
FR	347	1.175	367
IT	392	855	183
JP	366	49.904	761
SP	251	2.465	209
TR	280	176	26
ZH	982	22.536	407

whose difficulty is comparable across all corpora, we create a balanced subset for each corpus that contains the same number of documents for each language. This number is determined by the language with the fewest documents for each corpus (cf. Table 2). We randomly sample this number of documents from each other language. Table 1 shows the final statistics.

ICLE. The ICLE, version 2 (Granger et al. 2009), is the oldest and best-researched NLI corpus, a collection of essays written by students with a high intermediate to advanced level of English. The corpus originally comprises 16 different L1s.

TOEFL11. The TOEFL11 corpus (Blanchard et al. 2013) consists of texts that learners of English with mixed proficiency and 11 different native backgrounds wrote in response to prompts during TOEFL exams. The corpus was created as an alternative to the ICLE that is larger and more varied in subjects, but still well-controlled. It is balanced, consisting of 1100 documents per L1 class.

Lang-8. Lang8, introduced in (Brooke and Hirst 2011) is a web-scraped version of the Lang-8 website[1] where learners of various languages post texts for correction by native speakers. The collected documents are written by learners of English and span over 60 different native languages. Although Lang-8 counts as a learner corpus, it is relatively unfiltered and uncontrolled, showing artifacts similar to social media documents.

ACL-NLI. We adapted an approach proposed by Lamkiewicz (2014) to extract a novel dataset, ACL-NLI, from the 2012 release of the ACL Anthology Network Corpus (AAN, Radev et al. 2013). The AAN covers over 25,000 papers that appeared in conferences, workshops and journals sponsored by the Association for Computational Linguistics, dating back to the 1960s. While the AAN encodes author and e-mail information for each document explicitly as metadata entries, naturally the metadata does not encode the authors' L1. We worked on the assumption that a document could be assigned to an L1 category if and only if the email addresses of *all* of its authors belonged to a top-level country domain associated with the L1. While this heuristic would fail for countries with a high influx of foreign researchers, like the UK, and for countries which do not use a

1 `http://www.cs.toronto.edu/~jbrooke/Lang8.zip`

geographical top-level domain (like the US), it represents a reasonable, precision-oriented approach for the countries in our sample such as Turkey and Japan. Of course, it does not provide any guarantees, but in our manual evaluation of a small sample, we found its precision to be >95%.

This procedure yielded between 26 (Turkish) and 761 (Japanese) documents. To obtain a more satisfactory minimal number of documents per L1, we extended the set of documents for Turkish by specifying a small number of Turkish researchers working abroad. We finally obtained 54 documents for Turkish and used this number as the category size in ACL-NLI. Finally, we apply conservative preprocessing to each document. We removed title and headers, acknowledgments, and the bibliography, since these often contain hints regarding the authors' home country and L1.[2]

3. Modeling NLI Across Text Types with Domain Adaptation

We can now ask various research questions about our novel ACL-NLI corpus and its relationship to learner corpora. A first research question concerns its difficulty. Previous work on learner corpora has generally been able to achieve accuracies of 90% and above for various feature sets (cf. Section 1). It is not clear *a priori* what to expect for our novel ACL-NLI corpus. There are two reasons why NLI might be more difficult on ACL-NLI: (a) most authors in ACL-NLI have a better working knowledge of written English than typical learners and may avoid "typical" learner mistakes; (b) due to the conventions of scientific texts discussed in Section 1, structure and vocabulary are highly constrained and standardized. At the same time, the scientific articles are substantially longer than typical documents from learner corpora (cf. Table 1), which may make NLI easier.

A second question is how well NLI models can generalize across corpora. As described in the previous section, the learner corpora span a range from ICLE at one end (a well-controlled dataset) to Lang-8 at the other (a fairly free dataset), with TOEFL assuming an intermediate position. Thus, we would expect that NLI models are not generalizable even across learner corpora without additional effort (compare the discussions in Brooke and Hirst (2011) and Bykh and Meurers (2012)). When we move to our new dataset ACL-NLI, with its idiosyncratic properties, we expect that generalization becomes much more difficult.

To empirically assess these questions, we perform experiments on the learner corpora and the ACL-NLI, employing two *domain adaptation* techniques to improve the generalization of models.[3] We assume the "classical" domain adaptation scenario where we have a relatively large corpus in one domain (TOEFL11, as a well-controlled learner corpus) and combine it with smaller corpora (portions of ICLE, Lang-8, and ACL-NLI) to achieve better results on these corpora.[4] We use the standard terminology of *source* for the (main) training domain and *target* for the testing domain.

2 The ACL-NLI dataset is available at `http://www.ims.uni-stuttgart.de/data/nli`.

3 We use the established term "domain adaptation" although in our case, the differences among the learner corpora are arguably differences in genre, and ACL-NLI differs from the learner text type in both domain and genre.

4 An alternative proposed by Brooke and Hirst (2012a) to improve performance is to take advantage of simple L1 texts which are typically plentiful.

3.1 Feature Augmentation

The first domain adaptation method that we consider is Daumé III's (2007) *feature augmentation* strategy. It maps an input feature vector onto duplicated versions of itself, where each copy corresponds to one domain, thus allowing feature weights to be learned per domain. For two domains, for example, each feature in $\mathcal{X} = \mathbb{R}^F$ is mapped onto three new features via the mappings $\Phi^s, \Phi^t : \mathcal{X} \rightarrow \check{\mathcal{X}}$. The new feature space is $\check{\mathcal{X}} = \mathbb{R}^{3F}$. The mapping for the source-specific version is carried out to augment the source-domain dataset: $\Phi^s(x) = \langle x, x, \mathbf{0} \rangle$ with $\mathbf{0} = \langle 0, 0, ..., 0 \rangle \in \mathbb{R}^F$. Analogously, the mapping for the target-domain data is: $\Phi^t(x) = \langle x, \mathbf{0}, x \rangle$. The first version of each feature, the "general" version, is identical to the original feature. The second version, the "source" version, is identical to the general version for all instances from the source domain, and zero for all instances from the target domain; vice versa for the third version, the "target" version. This method can be applied as a preprocessing step to any learning algorithm.

3.2 Marginalized Stacked Denoising Autoencoders

Glorot, Bordes, and Bengio (2011) propose Stacked Denoising Autoencoders (SDA) for domain adaptation, multi-layer networks that reconstruct input data from a corrupted input by learning intermediate (hidden) layers. The output of this unsupervised method is used as the input to a learning algorithm. The intuition is that the intermediate layers model the relevant properties of the input data without overfitting, providing robust features that generalize well across domains. Since SDA is effective but slow to train, Chen, Xu, and Weinberger (2012) develop a marginalized SDA (mSDA) which makes the model more efficient by marginalizing out the corruptions.

Formally, the input data $\mathbf{x}$ is partially and randomly corrupted into $\tilde{\mathbf{x}}$ according to a corruption probability p. A simple corruption method is to set some of the feature values to zero. The autoencoder learns a hidden representation $h(\tilde{\mathbf{x}})$ from which x is reconstructed: $g(h(\tilde{\mathbf{x}})) \approx \mathbf{x}$. The objective is to minimize the reconstruction error $\ell(\mathbf{x}, g(h(\tilde{\mathbf{x}})))$. This procedure is done in m iterations, each time using a different corruption. If there are many more features than data points, Chen, Xu, and Weinberger (2012) use the x most frequent features. The data D is sliced into $\frac{D}{x} = y$ partitions and mSDA is performed on each partition y_i by decoding $g(h(\mathbf{y}_i)) \approx \mathbf{x}$. The y results are averaged to obtain the new representation. The new learned intermediate layer units are concatenated with the original features to create the representation.

Note that autoencoders, including mSDA, are more general than feature augmentation: they can be applied not only to two related datasets in domain adaptation, but also to individual datasets, to obtain novel features. We make use of this possibility for meta-parameter optimization in Section 4.

4. Experimental Setup

Experimental Design. Table 3 shows the set of experiments that we perform. We consider TOEFL11 as our source corpus and ICLE, Lang-8, and ACL-NLI as target corpora. To achieve comparability among experiments, we split each target corpus into a training set (2/3) and a test set (1/3). In all cases (except SRC-only), we perform three-fold cross-validation.

The two top rows show baseline experiments. In the SRC-only setup, we train on TOEFL11 and test on the target test sets. TGT-only trains and tests on each target corpus. The two bottom rows show domain adaptation experiments. The *-large* experiments

Table 3
Experimental configurations. Every setup is applied to three domains (ICLE, Lang-8, ACL-NLI). The last two rows each aggregate three setups ("large target" and "small target", respectively).

Experimental setup	Training Data	Test Data
SRC-only	TOEFL11	ICLE (1/3)
	TOEFL11	Lang-8 (1/3)
	TOEFL11	ACL-NLI (1/3)
TGT-only	ICLE (2/3)	ICLE (1/3)
	Lang-8 (2/3)	Lang-8 (1/3)
	ACL-NLI (2/3)	ACL-NLI (1/3)
{CONCAT,FA,mSDA}-large	TOEFL11 + ICLE (2/3)	ICLE (1/3)
	TOEFL11 + Lang-8 (2/3)	Lang-8 (1/3)
	TOEFL11 + ACL-NLI (2/3)	ACL-NLI (1/3)
{CONCAT,FA,mSDA}-small	TOEFL11 + ICLE (1/3)	ICLE (1/3)
	TOEFL11 + Lang-8 (1/3)	Lang-8 (1/3)
	TOEFL11 + ACL-NLI (1/3)	ACL-NLI (1/3)

combine the training sets used in the SRC-only and TGT-only scenarios, using three combination methods (plain concatenation and the two domain adaptation methods from Section 3). The final set of experiments (*-small*) is parallel to *-large*, but uses just one third of the target corpora, to assess the influence of the amount of training data.

All models use binary lexical features consisting of recurring unigrams and bigrams as proposed by Bykh and Meurers (2012). This type of feature has proven effective on its own (Bykh and Meurers 2012; Brooke and Hirst 2012b) and since we focus on cross-text type modelling, we choose to not use structural features. An n-gram is recurring if it occurs in more than two documents of the same class. As multi-class classifier, we use the LIBLINEAR Support Vector Machine library (Chang et al. 2008) with default parameters.

Evaluation and Baselines. We report classification accuracy and test statistical significance using the Chi-squared test with Yates' continuity correction (Yates 1984). Since our corpora are balanced across classes by design, the frequency (and random) baselines are at 1/7 = 14.3%.

Hyperparameters. The mSDA approach has a number of hyperparameters that need to be fixed. We set the corruption probability $p = 0.9$ and the number of layers $l = 1$ in line with previous work. To determine a reasonable number of most frequent features x, we apply mSDA individually to each of our target corpora, using the same data configurations as the SRC-only experiments shown in Table 3. The results in Figure 1 indicate optimal performance in the range of 5000 to 6000 features. Using more features also increases the run time of the experiments. Since the parameters for mSDA are usually chosen to balance a tradeoff between performance and speed, we set x to 5000 for all experiments.

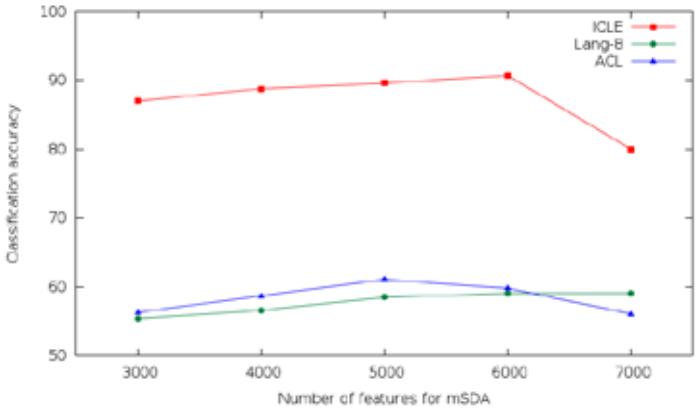

Figure 1
mSDA experiments to determine the best number of features x

Table 4
Classification accuracies. Bold indicates results not significantly different from best result for each test set ($p<0.05$). Significant improvements over result in previous row marked by asterisks (*: $p<0.05$, **: $p<0.01$, ***: $p<0.001$).

	Test sets		
Model	ICLE	Lang-8	ACL-NLI
SRC-only	79.5	57.7	49.5
TGT-only	96.1	77.1	**85.7**
CONCAT-large	94.4	80.0	75.1
FA-large	97.0***	84.1*	81.2
mSDA-large	**98.9****	**90.0****	**88.4****
CONCAT-small	92.5	75.5	68.8
FA-small	96.0***	77.9	74.6
mSDA-small	**98.6****	86.8***	**86.0****

5. Main Results

We start with a discussion of the quantitative results of our experiments, which are shown in Table 4.

5.1 Baseline Model Results

A comparison of the SRC-only and TGT-only experiments shows that training on TOEFL11 alone does not perform well on any of our three target corpora.[5] For ICLE and Lang-8, the difference is about 20 percentage points. For ACL-NLI, it is more than

5 Three-fold cross-validation on the TOEFL corpus with the same modeling choices yields an accuracy of 94%, indicating that the problem must be with the generalization and not with the model.

35%, confirming our hypotheses from Section 3 about the difficulty of generalizing to ACL-NLI. Interestingly, the TGT-only results also show that the NLI on the ACL-NLI corpus is not a priori more difficult: accuracy is 86%, which is 10% worse than the 96% that the model achieves on ICLE, but 9% better than the 77% on Lang-8.

In sum, these findings suggest that even within learner corpora, the datasets are sufficiently dissimilar that direct model reuse is problematic. In terms of absolute difficulty, therefore, the less well controlled Lang-8 behaves much more like the ACL-NLI dataset than like ICLE, which in turns clusters together with the TOEFL-11 dataset, the other "classical" learner corpus. This explains the good generalization results found by Bykh and Meurers (2012) but indicates that they may be restricted to "classical" learner corpora.

5.2 Domain Adaptation Results

The *-large* experiments establish an upper bound for the improvement that can be expected when source and target domain data is combined. Unsurprisingly, simple CONCATentation does not perform well, with degradation compared to TGT-only for ICLE (-2%) and ACL-NLI (-10%) and a small improvement for Lang-8 (+3%). Feature Augmentation yields some increases for ICLE and Lang-8, but not for ACL-NLI, which can be interpreted as improved generalization to similar, but not to more different, corpora. mSDA leads to the best results overall, leading to very high final accuracies between 88% (ACL-NLI) and 98% (ICLE). The concrete improvements over TGT-only are: ICLE +2.8%; Lang-8: +12.9%; ACL-NLI:+2.7%. This is the only method that improves on the in-domain ACL-NLI result, even though the improvement is not statistically significant, which we attribute to the small dataset size. We surmise that FA is handicapped by the relatively small sizes of Lang-8, and the very small size of the ACL-NLI corpus, which are "overpowered" by the large TOEFL-11 dataset. The *-small* experiments show an almost identical pattern to the *-large* experiments, but with overall lower numbers. The performance drops substantially for CONCAT and FA, but only somewhat for mSDA (ICLE: -0.3%, Lang-8: -3.2%, ACL-NLI: -2.4%; the difference is statistically significant only for Lang-8). This indicates that mSDA can already profit from relatively small target domain datasets.

In sum, these results show that the use of domain adaptation methods, in particular mSDA, to include TOEFL11 in models for smaller corpora, can improve the performance of NLI substantially. We do again see a major difference between learner corpora and ACL-NLI though: the improvement is statistically highly significant for both learner corpora (ICLE and Lang-8), but remains quite modest for ACL-NLI.

5.3 Confusion Matrices

As a more in-depth analysis, Figure 2 shows confusion matrices for the "simple" ICLE and the "difficult" ACL-NLI corpus in two experimental settings (TGT-only as baseline and mSDA-large as a better-generalizing model).

A comparison of the top and bottom rows shows the improvement from TGT-only to mSDA for both corpora, as a reduction in counts for the off-diagonal cells (that is, the misclassifications). It also shows clear differences between the two corpora. In the case of ICLE (left-hand column), many of the misclassifications in the TGT-only matrix correspond to linguistically plausible mistakes for documents from related languages (the Romance languages FR–SP–IT, and JP–ZH as a pair of East Asian languages). The mSDA misclassifications form an almost perfect subset of the TGT-

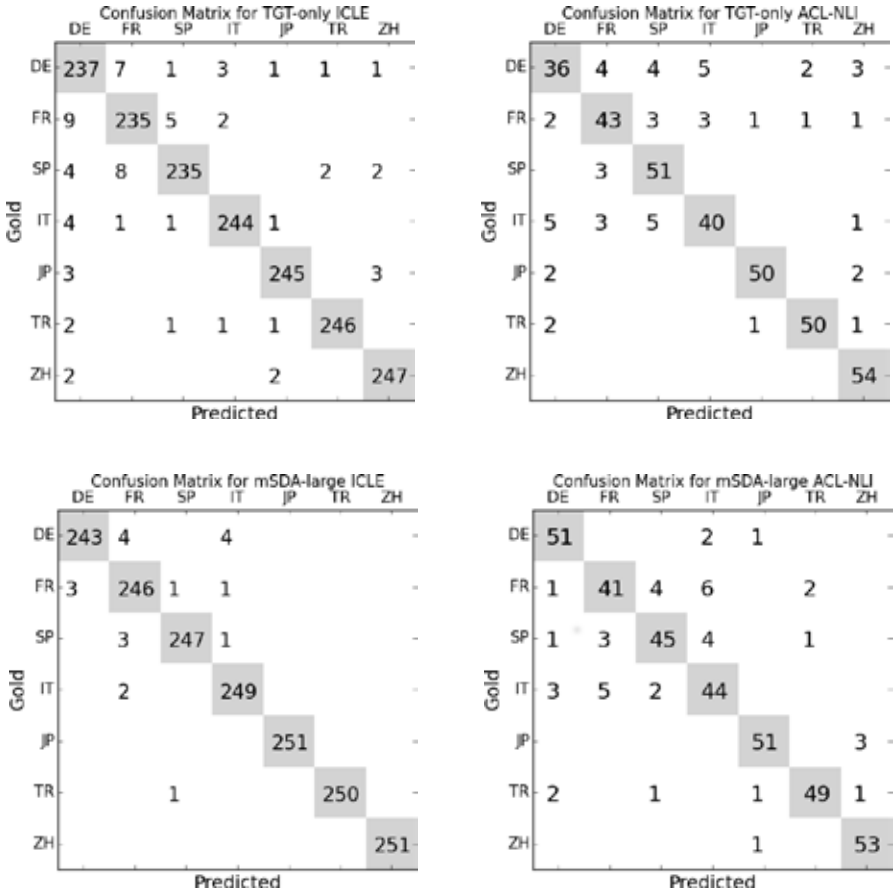

Figure 2
Confusion matrices for two corpora (ICLE,left and ACL-NLI, right) and two experimental settings (TGT-only, above and mSDA-large, below).

only ones, which is consistent with an interpretation of domain adaptation leading to a relatively conservative refinement of the model. Interestingly, it is exactly the linguistically plausible mistakes that are much reduced. All remaining problems are related to German documents (category pairs DE–FR and DE–IT). A closer look at the misclassified documents however reveals that this is in fact a topic-related phenomenon: the German documents tend to contain mentions of French culture (*Jean-Luc Godard, a famous French film-director*; *La révolution du language politique*) and vice versa (*In Germany some suggest [...] because it sounds like Mann (man)*).

In the case of ACL-NLI (right-hand column), the TGT-only misclassifications are much more varied and seem to be associated with particularly difficult languages (DE, FR, IT). An inspection of the documents indicates that many of these errors are "topical misclassification", i.e., documents on topics that are also worked on in another country. The situation flips for the ACL-NLI mSDA model: here, just as in the ICLE TGT-only model, many mistakes concern linguistically close language pairs (FR–SP–IT, JP–ZH). Thus, it seems as if domain adaptation has almost inverse effects on ICLE and ACL-NLI.

5.4 Feature Analysis

Table 5
Top 10 features of the Feature Augmentation FA-large model trained on TOEFL11+ICLE (source- and target-specific features) and TOEFL11+ACL-NLI (target-specific features).

Source-specific features (i.e., from TOEFL11) in TOEFL11+ICLE model						
Chinese	French	German	Italian	Japanese	Spanish	Turkish
with that	Indeed	, that	the	Japan	, etc	agree .
advertisements make	. Indeed	DIFFERENT	do because	is	specially	Turkey
that most	...	ABLE	particular	in Japan	option	dont
agree with	exemple	THIS	building the	disagree .	necesary	agree
that	concerned	IMPORTANT	Italy	two reasons	goals	enough
still	and give	MORE	, for	I disagree	not only	;
Take	France	, because	useful	Japanese	diferent	.
?	To conclude	ALL	the life	is because	, and	/
. Take	Indeed ,	OF	building	a products	activities	I agree
will	conclude	still	think that	Becouse	know about	can not
Target-specific features (i.e., from ICLE) in TOEFL11+ICLE model						
Chinese	French	German	Italian	Japanese	Spanish	Turkish
Kong	...	"	*	Japan	".	<92>
Hong Kong	may	-	children	Japanese	Spain	<92>s
Hong	television	:	society	don ’	"	<92>t
<R	the one	it	could	think	A	education
<	certainly	red	the *	I think	main	<93>
R >	free	that it	help	in Japan	think	the people
>	, we	be	words	?	would	example
R	"	you	’ s	so on	in which	universities
in Hong	But	next	too	to make	men	social
Kong .	than	friend	* is	the world	of this	don <92>
Target-specific features (i.e., from ACL-NLI) in TOEFL11+ACL-NLI model						
Chinese	French	German	Italian	Japanese	Spanish	Turkish
Chinese	French	German	a category	shows the	means of	Turkish
includes	of an	cf	implement	Figure 1	by means	give the
respectively	in French	the German	Given	Figure	es	suffixes
traditional	and their	allows us	22	used in	stored	for Turkish
of Chinese	lower	top	we obtain	cannot	trained	in Turkish
show that	tool	scores	. it	shows	is ,	inflectional
sides	the French	complete	availability of	In the	not so	the Turkish
testing	functions	us to	availability	Here	. es	Turkish ,
. Section	) with	see Figure	of category	1 shows	. Table	at a
translation ,	from a	we find	of (	which is	the system	building

To better understand the results of the previous subsection, we extract the top-ranked features of our models. Unfortunately, the features of the best model, mSDA, arising from hidden units of a neural network, are not human interpretable. Thus, we analyze the features of the Feature Augmentation FA-large models. As described in Section 3.1, FA uses one copy per feature for each domain. This enables us to qualitatively compare the top-ranked features learned on the source and target domains. Table 5 shows the features for TOEFL11, as learned as the source-specific features in the TOEFL+ICLE model[6] as well as the target-specific features for ICLE and ACL-NLI.

We first consider the TOEFL11 features (top table). As expected, we find a mixture of features that properly indicate the authors’ L1 “properly” via transfer of L1 language features and features that do so indirectly via their topics (“*I am from China*”). The topic

6 The source-specific features in the TOEFL+ACL-NLI model are essentially identical

features, however, are relatively underrepresented. We see one topic feature, the country name, for French, Italian, and Turkish, and three variations for Japanese. Italian is the only language where the features may be construed as belonging to a coherent topic ("life and society"). The language transfer features, on the other hand, cover a broad range of L1 influences. The French, Spanish and Japanese features include misspellings of cognates (*"exemple"*, *"necesary"*, *"becouse"*) while German authors are influenced by German punctuation rules which call for commas in front of embedded clauses (*", that"*, *", because"*). We also see lexical transfer expressed as the overuse of words that are more frequent in the L1 ("*concern*" for French). This may be a reflection of L1-specific register differences that were found to correlate with topics by Brooke and Hirst (2011): French writers in particular appear to adopt a formal, carefully argued style ("*To conclude*", "*Indeed*").

Moving on to the ICLE features (center table), we see a considerable number of artifacts related to encoding and representation issues. Beyond these, the topical aspects are more prominent for several languages. The top Chinese features are dominated by variations on *Hong Kong*, the Turkish features are consistent with a education-related topic, and Italian discusses similar society-related issues as in TOEFL11. Our interpretation is that an ICLE-only model accounts for topic bias at the expense of language transfer features, and thus has trouble dealing with documents from related languages that may not be clear-cut in terms of topics (the "linguistically plausible mistakes" from Section 5.3). The combined TOEFL11+ICLE model does a better job at distinguishing between related languages and therefore avoids these errors specifically.

The top features in the ACL-NLI corpus are also predominantly topic-related, with relatively little scope for language transfer, and we found considerably more varied and precise topic terms than in the ICLE. We find mentions of languages (*"Chinese"*, *"French"*), and many features reflect scientific terminology and preferred research topics. For example, Turkish researchers write about morphology (*"suffixes"*, *"inflectional"*), Spanish authors discuss Machine Learning (*"stored"*, *"trained"*, *"the system"*), and Chinese authors work on Machine Translation and Transliteration (*"traditional"*, *"translation"*). A clear case of L1 transfer is the German *"allows us"*, from *"erlaubt uns"*), which would be better translated as as *"enables us"*. For numerous features across languages, however, it is difficult to make a clear choice whether they indicate topics or language transfer. Notably for Italian, we find *"a category"*, *"implement"*, *"availability"*, *"we obtain"*, etc. Are these indicative of a preference for empirical study in Italy, or merely results of the (over-)use of particular collocations? While we cannot answer this at the moment, our analysis strongly indicates the ACL-NLI corpus can thus be considered to have an idiosyncratic form of topic bias — but one that is very different from the learner corpora.

Our conclusion from this analysis is that models trained solely on ACL-NLI will incorporate predominantly scientific topics. Thus, documents in languages with larger and less coherent research communities (like Germany and China) may be harder to model, which corresponds to our findings from Section 5.3. When the ACL-NLI data is combined with the TOEFL11 data, the combined model is quite different from the original ACL-NLI model — which explains the large difference between the confusion matrices of the two models — and assumes an intermediate position between modeling topics and language transfer similar to the ICLE-only model. Its stronger reliance on (imperfect) language transfer features can explain the language family patterns in the misclassifications.

Table 6
Classification results without country and language terms as features ("trivial topic indicators")

	ICLE	Lang-8	ACL-NLI
SRC-only	76.7	54.5	49.5
TGT-only	96.0	74.8	84.9
mSDA-large	98.7	88.2	86.5

5.5 Removing Trivial Topic Indicators

The previous subsection has identified two types of top-ranked features that can be seen as topic biases: features directly related to the authors' provenance (countries and languages), and content-related features (society and education in the case of ICLE and TOEFL11; computational linguistics in the case of ACL-NLI).

A minimal sanity check for the NLI models that we have discussed above is that they do not rely unduly on features of the first type, which would indicate an essentially trivial L1 classification. To test the importance of these features, we construct a stop word list for each L1 including the language name and country names (e.g. *Italian, Italy* for IT), including *Hong, Kong* for Chinese. We filter out all features that include these stop words.

Table 6 shows the results for the reduced feature set. They are somewhat, but not substantially lower than those in our previous experiments (cf. Table 4): we lose up to 3% in the SRC-only setting and up to 2% for TGT-only and mSDA-large. We conclude that mentions of country and language names do not unduly influence our results.

6. Conclusion

In this article, we have investigated the generalizability of models of Native Language Identification. We extended the previous discussion on generalizability among learner corpora to generalization to another corpus type, using domain adaptation methods. For this purpose, we constructed a NLI corpus of scientific documents from computational linguistics, ACL-NLI, sampled from the ACL Anthology Network.

An interesting picture emerges from our experiments: Native Language Identification on ACL-NLI, using just in-corpus data, is not necessarily more difficult; in fact, results are better than for Lang-8, a relatively free and social media-like learner corpus. However, ACL-NLI is "the odd one out" when it comes to generalization, since domain adaptation – at least with the methods we considered – only works well within text types, such as learner corpora (Bykh and Meurers 2012): Both ICLE and Lang-8 both profit from combination with TOEFL11, while the improvement for ACL-NLI is small. Our further analyses indicate that this difference is related to the question of what the models actually capture (language transfer or topic bias). The TOEFL11 model incorporates mostly language transfer features, which are arguably more generalizable than topic bias. Since the other, mostly smaller, learner corpora exhibit topic bias to varying extents, combining them with TOEFL11 can therefore improve their topic-tinged models with better language transfer features. The situation is different for ACL-NLI which exhibits a very different kind of topic bias: scientists from different countries work on different topics, which is reflected in the corpora. Evidently, relatively few of the TOEFL11 language transfer features are helpful for ACL-NLI – presumably because scientists

make different kinds of mistakes from learners, if only in terms of register. These results provide a more detailed insight into generalization patterns in NLI.

On the technical level, we find very good generalization performance for marginalized Stacked Denoising Autoencoders (mSDA), with the small downside that the resulting models are not human interpretable. Notably, the combination of TOEFL11 and half the ACL-NLI data in the mSDA-small experiment obtained the same performance as an in-domain model trained on the full ACL-NLI training set, thus reducing the data requirements by half even for a very different corpus. We believe that this is a promising result for modeling NLI on other low-resource domains.

With respect to future work, the most obvious avenue is a generalization of the present study to a more general feature set that includes longer n-grams or part-of-speech and structural features which yield the best state-of-the-art results (Tetreault, Blanchard, and Cahill 2013) and are arguably well suited to investigate genre-related differences (Petrenz and Webber 2011).

References

Baker, Mona. 1993. Corpus linguistics and translation studies. In *Text and Technology: In honour of John Sinclair*. John Benjamins, London, pages 233–250.

Blanchard, Daniel, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Marin Chodorow. 2013. TOEFL11: A Corpus of Non-Native English. *ETS Research Report Series*, 2013(2):1–15.

Brooke, Julian and Graeme Hirst. 2011. Native Language Detection with 'Cheap' Learner Corpora. In *Proceedings of the 2011 Conference on Learner Corpus Research*, Louvain-la-Neuve, Belgium.

Brooke, Julian and Graeme Hirst. 2012a. Measuring interlanguage: Native language identification with L1-influence metrics. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 779–784, Istanbul, Turkey.

Brooke, Julian and Graeme Hirst. 2012b. Robust, Lexicalized Native Language Identification. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 391–408, Mumbai, India.

Bykh, Serhiy and Detmar Meurers. 2012. Native Language Identification Using Recurring N-Grams – Investigating Abstraction and Domain Dependence. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 425–440, Mumbai, India.

Bykh, Serhiy and Detmar Meurers. 2014. Exploring syntactic features for native language identification: A variationist perspective on feature encoding and ensemble optimization. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1962–1973, Dublin, Ireland.

Cardinaletti, Anna and Giuliana Garzone, editors. 2005. *L'italiano delle traduzioni*. FrancoAngeli, Milan, Italy.

Chang, Kai-Wei, Cho-Jui Hsieh, Xiang-Rui Wang, Chih-Jen Lin, and Soeren Sonnenburg. 2008. Liblinear: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874.

Chen, Minmin, Zhixiang (Eddie) Xu, and Kilian Q. Weinberger. 2012. Marginalized Denoising Autoencoders for Domain Adaptation. In *Proceedings of the 29th International Conference on Machine Learning*, pages 767–774, Edinburgh, Scotland.

Cimino, Andrea, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2013. Linguistic profiling based on general-purpose features and native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 207–215, Atlanta, Georgia.

Daumé III, Hal. 2007. Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic.

Estival, Dominique, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. 2007. Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 263–272, Melbourne, Australia.

Galvão, Gabriela. 2009. Linguistic interference in translated academic texts: A case study of portuguese interference in abstracts translated into english. Bachelor's Thesis.

Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. 2011. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. In *Proceedings of the 28th International*

Conference on Machine Learning, pages 97–110, Bellevue, WA.
Granger, Sylviane, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English, Version 2*. Presses Universitaires de Louvain, Louvain-la-Neuve, Belgium.
Hyland, Ken. 2009. *Academic Discourse*. Continuum, London.
Lamkiewicz, Anna Maria. 2014. Automatische Erkennung der Muttersprache von L2-Englisch-Autoren. Magisterarbeit, Institut für Computerlinguistik, Neuphilologische Fakultät, Ruprecht-Karls-Universität Heidelberg.
Malmasi, Shervin and Mark Dras. 2015. Large-scale native language identification with cross-corpus evaluation. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 1403–1409, Denver, Colorado.
McEnery, Tony and Paul Baker. 2003. Corpora, translation and multilingual computing. In Federico Zanettin, Silvia Bernardini, and Dominic Stewart, editors, *Corpora in Translator Education*. St. Jerome Publishing, pages 89–102.
Odlin, Terence. 1989. *Language Transfer: Cross-linguistic influence in language learning*. Cambridge University Press.
Olohan, Maeve and Myriam Salama-Carr, editors. 2011. *Special issue on Science in Translation*, volume 17(2) of *The Translator*. St. Jerome Publishing.
Perkins, Ria. 2015. Native language identification (NLID) for forensic authorship analysis of weblogs. In Maurice Dawson and Marwan Omar, editors, *New Threats and Countermeasures in Digital Crime and Cyber Terrorism*. ISI Global, pages 213–234.
Petrenz, Philipp and Bonnie Webber. 2011. Stable classification of text genres. *Computational Linguistics*, 37(2):385–393.
Radev, Dragomir R., Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The ACL anthology network corpus. *Language Resources and Evaluation*, 47(4):919–944.
Swales, John. 2004. *Research Genres: Explorations and Applications*. Cambridge University Press.
Tetreault, Joel, Daniel Blanchard, and Aoife Cahill. 2013. A Report on the First Native Language Identification Shared Task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–57, Atlanta, Georgia.
Tetreault, Joel, Daniel Blanchard, Aoife Cahill, and Martin Chodorow. 2012. Native Tongues, Lost and Found: Resources and Empirical Evaluations in Native Language Identification. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 2585–2602, Mumbai, India.
Teufel, Simone and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.
Truffaut, Louis. 1997. *Traducteur tu seras*. Éditions du Hazard, Brussels, Belgium.
Tsur, Oren and Ari Rappoport. 2007. Using Classifier Features for Studying the Effect of Native Language on the Choice of Written Second Language Words. In *Proceedings of the ACL Workshop on Cognitive Aspects of Computational Linguistics*, pages 9–16, Prague, Czech Republic.
Wong, Sze-Meng Jojo and Mark Dras. 2011. Exploiting parse structures for native language identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1600–1610, Edinburgh, Scotland.
Yates, Frank. 1984. Tests of Significance for 2x2 Contigency Tables. *Journal of the Royal Statistical Society Series A*, 147 (3):426–463.