Cross-Lingual Lexico-Semantic Transfer in Language Learning

Ekaterina Kochmar
The ALTA Institute
University of Cambridge
ek358@cam.ac.uk

Ekaterina Shutova
Computer Laboratory
University of Cambridge
es407@cam.ac.uk

Abstract

Lexico-semantic knowledge of our native language provides an initial foundation for second language learning. In this paper, we investigate whether and to what extent the lexico-semantic models of the native language (L1) are transferred to the second language (L2). Specifically, we focus on the problem of lexical choice and investigate it in the context of three typologically diverse languages: Russian, Spanish and English. We show that a statistical semantic model learned from L1 data improves automatic error detection in L2 for the speakers of the respective L1. Finally, we investigate whether the semantic model learned from a particular L1 is portable to other, typologically related languages.

1 Introduction

Lexico-semantic knowledge of our native language is one of the factors that underlie our ability to communicate and reason about the world. It is also the knowledge that guides us in the process of second language learning. Lexico-semantic variation across languages (Bach and Chao, 2008) makes lexical choice a challenging task for second language learners (Odlin, 1989). For instance, the meaning of the English expression *pull the trigger* is realised as **push the trigger* in Russian and Spanish, possibly leading to errors of lexical choice by Russian and Spanish speakers learning English. Our native language (L1) plays an essential role in the process of lexical choice. When choosing between several linguistic realisations in L2, non-native speakers may rely on the lexico-semantic information from L1 and select a translational equivalent that they deem to match their communicative intent best. For example, Russian speakers **do exceptions* and *offers* instead of *making* them, and **find decisions* instead of *finding solutions*, since in Russian *do* and *make* have a single translational equivalent (*delat'*), and so do *decision* and *solution* (*resheniye*). As a result, non-native speakers who tend to fall back to their L1 translate phrases word-for-word, violating English lexico-semantic conventions.

The effect of L1 interference on lexical choice in L2 has been pointed out in a number of studies (Chang et al., 2008; Rozovskaya, 2010; Rozovskaya, 2011; Dahlmeier and Ng, 2011). Some of these studies also demonstrated that using L1-specific properties, such as the error patterns of speakers of a given L1 or L1-induced paraphrases, improves the performance of automatic error correction in non-native writing. However, neither of the approaches has constructed a semantic model from L1 data and systematically studied the effects of its transfer onto L2. In addition, most previous work has focused on error correction, bypassing the task of error detection for lexical choice. Lexical choice is one of the most challenging tasks for both non-native speakers and automated error detection and correction (EDC) systems. The results of the most recent shared task on EDC, which spanned all error types including lexical choice, show that most teams either did not propose any algorithms for this type of errors or did not perform well on them (Ng, 2014).

In this paper, we experimentally investigate the influence of L1 on lexical choice in L2 and whether lexico-semantic models from L1 are transferred to L2 during language learning. For this purpose, we induce L1 and L2 semantic models from corpus statistics in each language independently, and then use the discrepancies between the two models to identify errors of lexical choice. We focus on two types of *verb–noun* combinations, VERB–DIRECT_OBJECT (*dobj*) and

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 974–983,
Berlin, Germany, August 7-12, 2016.

SUBJECT–VERB (*subj*), and consider two widely spoken L1s from different language families – Russian and Spanish. We conduct our experiments using the Cambridge Learner Corpus (Nicholls, 2003), containing writing samples of non-native speakers of English. Spanish speakers account for around 24.6% of the non-native speakers represented in this corpus and Russian speakers for 4%.

Our experiments test two hypotheses: (1) that L1 effects in the lexical choice in L2 reveal themselves in the difference of the word association strength in the L1 and L2; and (2) that L1 lexico-semantic models are portable to other, typologically related languages. To the best of our knowledge, our paper is the first one to experimentally investigate these questions. Our results demonstrate that L1-induced information improves automatic error detection for lexical choice, confirming the hypothesis that L1 speakers rely on semantic knowledge from their native language during L2 learning. We test the second hypothesis by verifying that Russian speakers exhibit similar trends in errors with the speakers of other Slavic languages, and Spanish speakers with the speakers of other Romance languages. We find that the L1-induced information from Russian and Spanish is effective in assessing lexical choice of the speakers of other languages for both language groups.

2 Related work

2.1 Error detection in content words

Early approaches to collocation error detection relied on manually created databases of correct and incorrect word combinations (Shei and Pain, 2000; Wible et al., 2003; Chang et al., 2008). Constructing such databases is expensive and time-consuming, and therefore, more recent research turned to the use of machine learning techniques.

Leacock et al. (2014) note that most approaches to detection and correction of collocation errors compare the writer's word choice to the set of alternatives using association strength measures and choose the combination with the highest score, reporting an error if this combination does not coincide with the original choice (Futagi et al., 2008; Östling and Knutsson, 2009; Liu et al., 2009). This strategy is expensive as it relies on comparison with a set of alternatives, limited in capacity as it depends on the quality of the alternatives generated and circular as the detection cannot be performed independently of the correction. Our approach alleviates these problems, since error detection depends on the original combination only.

Some previous approaches focused on correction only (Dahlmeier and Ng, 2011; Kochmar and Briscoe, 2015), and although they show promising results, they have not attempted to perform error detection in lexical choice. Kochmar and Briscoe (2014) focus on error detection, but their system addresses adjective–noun combinations and does not use L1-induced information.

2.2 L1 factors in L2 writing

The influence of an L1 on lexical choice in L2 and the resulting errors have been previously studied (Chang et al., 2008; Östling and Knutsson, 2009; Dahlmeier and Ng, 2011). These works focus on errors in particular L1s and use the translational equivalents directly to improve candidate selection and quality of corrections. Dahlmeier and Ng (2011) show that L1-induced paraphrases outperform approaches based on edit distance, homophones, and WordNet synonyms in selecting the appropriate corrections.

Rozovskaya and Roth (2010) show that an error correction system for prepositions benefits from restricting the set of possible corrections to those observed in the non-native data. Rozovskaya and Roth (2011) further demonstrate that the models perform better when they use knowledge about error patterns of the non-native writers. According to their results, an error correction algorithm that relies on a set of priors dependent on the writer's preposition and the writer's L1 outperforms other methods. Madnani et al. (2008) show promising results in whole-sentence grammatical error correction using round-trip translations from Google Translate via 8 different pivot languages.

The results of these studies suggest that L1 is a valuable source of information in EDC. However, all these works use isolated translational equivalents and focus on error correction only. In contrast, we construct holistic semantic models of L1 from L1 corpora and use these models to perform the more challenging task of error detection.

3 Data

We first use large monolingual corpora in Spanish, Russian and English to build word association models for each of the languages. We then apply the resulting models for error detection in the English learner data.

3.1 L1 Data

Spanish data The Spanish data was extracted from the Spanish Gigaword corpus (Mendonca et al., 2011), a one billion-word collection of news articles in Spanish. The corpus was parsed using the Spanish Malt parser (Nivre et al., 2007; Ballesteros et al., 2010). We extracted VERB–SUBJECT and VERB–DIRECT_OBJECT relations from the output of the parser, which we then used to build an L1 word association model for Spanish.

Russian data The Russian data was extracted from the RU-WaC corpus (Sharoff, 2006), a two billion-word representative collection of texts from the Russian Web. The corpus was parsed using Malt dependency parser for Russian (Sharoff and Nivre, 2011), and the VERB–SUBJECT and VERB–DIRECT_OBJECT relations were extracted from the parser output to create an L1 word association model for Russian.

Dictionaries and translation Once the L1 word associations have been computed for the verb–noun pairs, we identify possible translations for verbs and nouns (in each pair) in isolation, as a language learner might do. To create the translation dictionaries, we extracted translations from the English–Spanish and English–Russian editions of Wiktionary, both from the translation sections and the gloss sections if the latter contained single words as glosses. We focus on verb–noun pairs, therefore multi-word expressions were universally removed. We added inverse translations for every original translation. We then created separate translation dictionaries for each language and part-of-speech tag combination from the resulting collection of translations.

3.2 L2 data

To build the English word association model, we have used a combination of the British National Corpus (Burnard, 2007) and the UKWaC (Baroni et al., 2009). The corpora were parsed by the RASP parser (Briscoe et al., 2006) and VERB–SUBJECT and VERB–DIRECT_OBJECT relations were extracted from the parser output. Since the UKWaC is a Web corpus, we assume that the data contains a certain amount of noise, e.g. typographical errors, slang and non-words. We filter these out by checking that the verbs and nouns in the extracted relations are included in WordNet (Miller, 1995) with the appropriate part of speech.

3.3 Learner data

To extract the verb–noun combinations that have been used by non-native speakers in practice, we use the *Cambridge Learner Corpus* (CLC), which is a 52.5 million-word corpus of learner English collected by Cambridge University Press and Cambridge English Language Assessment since 1993 (Nicholls, 2003). It comprises English examination scripts written by learners of English with 148 different L1s, ranging across multiple examinations and covering all levels of language proficiency. A 25.5 million-word component of the CLC has been manually error-annotated.

We have preprocessed the CLC with the RASP parser (Briscoe et al., 2006), as it is robust when applied to ungrammatical sentences. We have then extracted all *dobj* and *subj* combinations: in total, we have extracted $187,109$ *dobj* and $225,716$ *subj* combinations. We have used the CLC error annotation to split the data into correct combinations and errors. We note that some verb–noun combinations are annotated both as being correct and as errors, depending on their wider context of use. To ensure that the annotation we use in our experiments is reliable and not context-dependent, we have empirically set a threshold to filter out ambiguously annotated instances. The set of correct word combinations includes only those word pairs that are used correctly in at least 70% of the cases they occur in the CLC; the set of errors includes only those that are used incorrectly at least 70% of the time.

3.4 Experimental datasets

We split the annotated CLC data by language and relation type. Table 1 presents the statistics on the datasets collected.[1] We extract the verb–noun combinations from the CLC texts written by native speakers of Russian (RU) and Spanish (ES) to test our first hypothesis, as well as by speakers of ALL L1s in the CLC to test our second hypothesis. We then filter the extracted relations using the translated verb–noun pairs from Russian and Spanish corpora.

We note that Russian and Spanish have comparable number of word combinations in L1-specific subsets – $10K$-$12K$ for *dobj* and *subj* combinations – and comparable error rates (ERR). We also note that the error rates in the *dobj* sub-

[1]The data is available at `http://www.cl.cam.ac.uk/~ek358/cross-ling-data.html`

Source	CLC	Total	ERR (%)	verbs	nouns
RU_{dobj}	RU	11,184	12.55	786	1,918
	ALL	62,923	14.02	1,387	4,168
RU_{subj}	RU	10,417	7.90	734	1,775
	ALL	63,649	9.49	1,403	4,374
ES_{dobj}	ES	11,959	14.66	705	1,926
	ALL	32,966	15.17	1,072	2,928
ES_{subj}	ES	9,899	8.09	573	1,733
	ALL	26,766	9.42	877	2,762

Table 1: Statistics on the datasets collected.

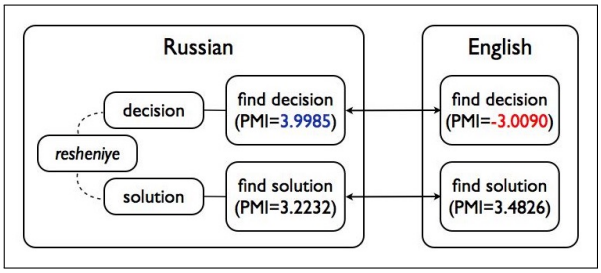

Figure 1: Russian to English interface for **find decision*.

sets are higher than in *subj* subsets, presumably, because VERB–SUBJECT combinations allow for more flexibility in lexical choice. We find a large number of translated word combinations in other L1s, and it is interesting to note that the error rates are higher across multiple languages than in the same L1s, which corroborates our second hypothesis that the lexico-semantic models from L1s transfer to L2. The last two columns of Table 1 show how diverse our datasets are in terms of verbs and nouns used in the constructions: for example, RU_{dobj} subset contains combinations with 786 different verbs and 1,918 different nouns.

4 Methods

Our approach to detecting lexico-semantic transfer errors relies on the intuition that a mismatch between the lexico-semantic models in two languages reveals itself in the difference in word association scores. We argue that a high association score of a *verb–noun* combination in L1 shows that it is a collocation in L1, but low association score of its translational equivalent in L2 signals an error in L2 stemming from the lexico-semantic transfer. Following previous research (Baldwin and Kim, 2010), we measure the strength of verb–noun association using pointwise mutual information (PMI). Figure 1 illustrates this intuition. In Russian, both **find decision* vs. *find solution* have a high PMI score. However, in English the latter has a high PMI while the former has a negative PMI. We expect such a discrepancy in word association to be an indicator of error of lexical choice, driven by the L1 semantics.

We treat the task of lexico-semantic transfer error detection as a binary classification problem and train a classifier for this task. The classifier uses a combination of L1 and L2 semantic features. If our hypothesis holds, we expect to see an improvement in the classifier's performance when adding L1 semantic features.

4.1 L2 lexico-semantic features

We experiment with two types of L2 features: lexico-semantic features and semantic vector space features.

Lexico-semantic features include:

- `pmi` in L2: we estimate the association strength between the noun and verb using the combined BNC and UKWaC corpus;
- `verb` and `noun`: the identity of the verb and the noun in the pair, encoded in a numerical form in the range of $(0, 1)$. The motivation behind that step is that certain words are more error-prone than others and converting them into numerical features helps the classifier to use this information.

Semantic vector space features Kochmar and Briscoe (2014) obtained state-of-the-art results in error detection by using the semantic component of the content word combinations. We reimplement these features and test their impact on our task. We extracted the noun and verb vectors from the publicly available `word2vec` dataset of word embeddings for 3 million words and phrases.[2] The 300-dimensional vectors have been trained on a part of Google News dataset (about 100 billion words) using `word2vec` (Mikolov et al., 2013). The *dobj* and *subj* vectors are then built using element-wise addition on the vectors (Mitchell and Lapata, 2008; Mikolov et al., 2013; Kochmar and Briscoe, 2014).

Once the compositional vectors are created, the method relies on the idea that correct combinations can be distinguished from the erroneous ones by certain vector properties (Vecchi et al., 2011; Kochmar and Briscoe, 2014). We implement a set of numerical features based on the following properties of the vectors:

[2] `code.google.com/archive/p/word2vec/`

- `length` of the additive (vn) vector
- $\texttt{cos}_{vn \wedge n}$ – cosine between the vn vector and the noun vector
- $\texttt{cos}_{vn \wedge v}$ – cosine between the vn vector and the verb vector
- $\texttt{dist}_{10}$ – distance to the 10 nearest neighbours of the vn vector
- `lex-overlap` – proportion of the 10 nearest neighbours of the vn vector containing the verb/noun
- `comp-overlap` – overlap between the 10 neighbours of the vn vector and 10 neighbours of the verb/noun vector
- $\texttt{cos}_{v \wedge n}$ – cosine between the verb and the noun vectors.

The 10 nearest neighbours are retrieved in the combined semantic space containing word embeddings and additive phrase vectors. All features, except for the last one, have been introduced in previous work and showed promising results (Vecchi et al., 2011; Kochmar and Briscoe, 2014). For example, it has been shown that the distance from the constructed word combination vector to its nearest neighbours is one of the discriminative features of the error detection classifier. Manual inspection of the vectors and nearest neighbours shows that the closest neighbour to **find decision* is *see decision* with the similarity of 0.8735 while the closest one to *find solution* is *discover solution* with the similarity of 0.9048.

We implement an additional $\text{cos}_{v \wedge n}$ feature based on the intuition that the distance between the verb and noun vectors themselves may indicate a semantic mismatch and thus help in detecting lexical choice errors.

4.2 L1 lexico-semantic features

We first quantified the strength of association between the L1 verbs and nouns in the original L1 data, using PMI. We then generated a set of possible translations for each verb–noun pair in L1 using the translation dictionaries. Each verb–noun pair in the CLC was then mapped to one of the translated L1 pairs and its L1 features. We used the following L1 features in classification:

- `pmi` in L1: we estimate the strength of association on the original L1 corpora;
- `difference` between the PMI of the verb–noun pair in L1 and in L2.

4.3 Classification

Classifier settings We treat the task as a binary classification problem and apply a linear SVM classifier using `scikit-learn LinearSVC` implementation.[3] The error rates in Table 1 show that we are dealing with a two-class problem where one class (*correct* word combinations) significantly outnumbers the other class (*errors*) by up to 11:1 (on RU_{subj}). To address the problem of class imbalance, we use subsampling: we randomly split the set of correct word combinations in n samples keeping the majority class baseline under 0.60, and run n experiments over the samples. We apply 10-fold cross-validation within each sample. The results reported in the following sections are averaged across the samples for each dataset.

Evaluation The goal of the classifier is to detect errors, therefore we primarily focus on its performance on the *error* class and, in addition to accuracy, report *precision* (P), *recall* (R) and F_1 on this class. Previous studies (Nagata and Nakatani, 2010) suggest that systems with high precision in detecting errors are more helpful for L2 learning than systems with high recall as non-native speakers find misidentified errors very misleading. In line with this research, we focus on maximising precision on the *error* class.

Baseline We compare the performance of our different feature sets to the baseline classifier which uses L2 co-occurrence frequency of the verb and noun in the pair as a single feature. Frequency sets a competitive baseline as it is often judged to be the measure of acceptability of an expression and many previous works relied on the frequency of occurrence as an evidence of acceptability (Shei and Pain, 2000; Futagi et al., 2008).

5 Experimental Results

To test our hypothesis that lexico-semantic models are transferred from L1 to L2, we first run the set of experiments on the L1 subsets of the CLC data, that is $\text{RU} \rightarrow \text{RU}_{CLC}$ and $\text{ES} \rightarrow \text{ES}_{CLC}$, where the left-hand side of the notation denotes the lexico-semantic model and the right-hand side the L1 of the speakers that produced the word pairs extracted from the CLC. We incrementally add the features, starting with the set of lexico-semantic

[3] `scikit-learn.org/`

L1	Features	Acc	$\mathbf{P}_e$	R_e	F_{1_e}
RU_{dobj}	baseline	55.68	47.77	61.44	53.55
	pmi_{En}	64.74	59.76	47.55	52.96
	+verb	64.79	59.87	47.56	53.01
RU_{subj}	baseline	54.48	46.30	63.96	53.17
	pmi_{En}	67.02	58.86	62.74	60.74
	+verb	67.64	59.84	62.17	60.98
ES_{dobj}	baseline	56.74	52.25	74.44	61.36
	pmi_{En}	64.28	61.75	59.55	60.63
	+verb	64.34	61.80	59.67	60.71
ES_{subj}	baseline	54.45	46.71	70.31	56.00
	pmi_{En}	69.22	61.35	68.83	64.87
	+verb	69.51	61.79	68.58	65.00

Table 2: System performance (in %) using L2 lexico-semantic features, L1 $\rightarrow$ L1$_{CLC}$.

features in L2 that are readily available without reference to the L1, and later adding L1 semantic features, and measure their contribution.

5.1 L2 lexico-semantic features

The first system configuration we experiment with uses the set of lexico-semantic features from L2. Table 2 reports the results. Our experiments show that a classifier that uses L2 PMI ($\mathtt{pmi}_{En}$) as a single feature performs with relatively high accuracy: on all four datasets it outperforms the baseline classifier achieving an increase from 7.54% (on ES_{dobj}) up to 14.77% (on ES_{subj}) in accuracy.

Adding the `noun` as a feature decreases performance of the classifier and we do not further use this feature. The `verb` used as an additional feature consistently improves classifier performance.

5.2 L2 semantic vector space features

Next, we test the combination of the semantic vector space features (`sem`) and combine them with two L2 lexico-semantic features including $\mathtt{pmi}_{En}$ and `verb` (denoted as $\mathtt{ft}_{En}$ hereafter for brevity). Table 3 reports the results.

We note that the semantic vector space features on their own yield precision of $50\% - 52\%$ on the *error* class in *dobj* combinations and lower than 50% on *subj* combinations. This suggests that the classifier misidentifies correct combinations as errors more frequently than it correctly detects errors. Moreover, recall of this system configuration is also low on all datasets. Adding the semantic vector space features to the other L2 semantic features, however, improves the performance, as shown in Table 3. As both groups of features refer to the phenomena in L2, the results suggest that they complement each other.

L1	Features	Acc	$\mathbf{P}_e$	R_e	F_{1_e}
RU_{dobj}	sem	58.36	50.72	6.98	12.22
	+ft_{En}	65.90	58.64	62.18	60.35
RU_{subj}	sem	58.62	36.07	3.40	6.12
	+ft_{En}	68.37	60.05	66.48	63.07
ES_{dobj}	sem	54.51	52.01	20.78	29.48
	+ft_{En}	66.87	63.36	67.08	65.16
ES_{subj}	sem	58.63	49.37	9.27	15.47
	+ft_{En}	70.75	62.21	74.31	67.72

Table 3: System performance (in %) using a combination of L2 semantic features, L1 $\rightarrow$ L1$_{CLC}$.

L1	Features	Acc	$\mathbf{P}_e$	R_e	F_{1_e}
RU_{dobj}	ft_{En}	64.79	59.87	47.56	53.01
	+pmi_{L1}	66.05	58.74	62.72	60.67
RU_{subj}	ft_{En}	67.64	59.88	62.17	60.98
	+pmi_{L1}	68.68	62.10	69.61	64.38
ES_{dobj}	ft_{En}	64.34	61.80	59.67	60.71
	+pmi_{L1}	66.89	63.01	68.61	65.68
ES_{subj}	ft_{En}	69.51	61.79	68.58	65.00
	+pmi_{L1}	71.19	62.10	77.66	69.00

Table 4: System performance (in %) using L1 and L2 lexico-semantic features, L1 $\rightarrow$ L1$_{CLC}$.

5.3 L1 lexico-semantic features

Finally, we add the L1 lexico-semantic features to the well-performing L2 features (`pmi` and `verb`). The combination of L1 lexico-semantic features with the L2 lexico-semantic and semantic vector space features achieves lower results, therefore we do not report them here. The use of L1 `pmi` improves both the accuracy and the F-score of the *error* class (see Table 4). For the ease of comparison, we also include the results obtained using a combination of L1 lexico-semantic features (denoted $\mathtt{ft}_{En}$). The addition of the explicit `difference` feature between the two PMIs has not yielded further improvement. This is likely to be due to the fact that the classifier already implicitly captures the knowledge of this difference in the form of individual L1 and L2 PMIs.

We note that the system using a combination of L1 and L2 lexico-semantic features gains an absolute improvement in accuracy from 1.04% for RU_{subj} to 2.55% on ES_{dobj}. The performance on the *error* class improves in all but one case (P_e on RU_{dobj}), with an absolute increase in F_1 up to 7.66%. The system has both a higher coverage in error detection (a rise in recall) and a higher precision. The improvement in performance across all four datasets is statistically significant at 0.05 level. These results demonstrate the effect of lexico-semantic model transfer from L1 to L2.

6 Effect on different L1s

Next, we test our second hypothesis that a lexico-semantic model from one L1 is portable across several L1s, in particular, typologically related ones. We first experiment with the data representing all L1s in the CLC and then with the data representing a specific language group. We compare the performance of the baseline system using verb–noun co-occurrence frequency as a single feature, the system that uses L2 semantic features only and the system that combines both L2 and L1 semantic features.

6.1 Experiments on all L1s

Table 1 shows that using the translated verb–noun combinations from our L1s (RU and ES) we are able to find a large amount of both correct and erroneous combinations in different L1s in the CLC including RU and ES (see ALL). This gives us an initial confirmation that the lexico-semantic models may be shared across multiple languages.

We then experiment with error detection across all L1s represented in the CLC. The results are shown in Table 5. The baseline system achieves similar performance on RU $\rightarrow$ ALL$_{CLC}$ as on RU $\rightarrow$ RU$_{CLC}$, and better performance on ES $\rightarrow$ ALL$_{CLC}$ than on ES $\rightarrow$ ES$_{CLC}$. The results obtained with the L2 lexico-semantic features are also comparable: the system achieves an absolute increase in accuracy of up to 9.86% for the model transferred from RU$_{subj}$, reaching an accuracy of around $65 - 66\%$ with balanced performance in terms of precision and recall on *errors*.

When the L1 lexico-semantic features are added to the model, we observe an absolute increase in the accuracy ranging from 0.57% (for RU$_{subj}$) to 1.43% (for ES$_{dobj}$). The Spanish lexico-semantic model has a higher positive effect on all measures, including precision on the *error* class. Although the addition of the L1 lexico-semantic features does not have a significant effect on the accuracy and precision, the system achieves an absolute improvement in recall of up to 12.71% (on RU$_{dobj}$). That is, the system that uses L1 lexico-semantic features is able to find more errors in the data originating with a set of different L1s. Generally, the results of the Spanish model are more stable and comparable to the results in the previous Section, which may be explained by the fact that Spanish is more well-represented in the CLC.

L1	Features	Acc	P_e	R_e	F_{1_e}
RU$_{dobj}$	baseline	55.13	50.17	72.14	58.99
	ft$_{En}$	63.58	59.73	57.98	58.85
	+pmi$_{L1}$	64.60	58.81	70.69	64.20
RU$_{subj}$	baseline	54.56	47.95	71.10	56.71
	ft$_{En}$	64.42	57.27	62.64	59.83
	+pmi$_{L1}$	64.99	57.24	68.17	62.21
ES$_{dobj}$	baseline	59.35	55.38	71.87	62.51
	ft$_{En}$	64.32	61.89	63.47	62.67
	+pmi$_{L1}$	65.75	61.90	71.37	66.30
ES$_{subj}$	baseline	58.34	50.90	66.97	57.48
	ft$_{En}$	65.57	58.32	64.09	61.06
	+pmi$_{L1}$	66.54	58.80	68.72	63.36

Table 5: System performance (in %) using L1 and L2 lexico-semantic features, L1 $\rightarrow$ all L1s.

6.2 Experiments on related L1s

The results on ALL L1s confirm our expectations: since we have extracted verb–noun combinations that originate with two particular L1s from the set of all different L1s in the CLC, and then used the L1 lexico-semantic features, the system is able to identify more errors thus we observe an improvement in recall. The precision, however, does not improve, possibly because the set of errors in ALL L1s is different from that in the two L1s we rely on to build the lexico-semantic models. The final question that we investigate is whether the lexico-semantic models of our L1s are directly portable to typologically related languages. If this is the case, we expect to see an effect on the precision of the classifier as well as on the recall.

We experiment with the following groups of related languages ordered by the number of verb–noun pairs we found in the CLC data:

- `Ru group`: Russian, Polish, Czech, Slovak, Serbian, Croatian, Bulgarian, Slovene;
- `Es group`: Spanish, Italian, Portuguese, French, Catalan, Romanian, Romansch.

In addition to investigating the effect of the L1 lexico-semantic model on the whole language group, we also consider its effects on individual languages. We chose Polish for the RU model, and Italian for the ES model as these two languages have the most data representing their native speakers in the CLC. Table 6 shows the number of verb–noun combinations and error rates for the language groups and these individual languages.

The results are presented in Tables 7 and 8. They exhibit similar trends in the change of the system performance on L1 $\rightarrow$ L1_GROUP as we

Source	Targets	Total	ERR
RU$_{dobj}$	Slavic	18,721	9.19
	Polish	11,327	8.16
RU$_{subj}$	Slavic	18,511	6.80
	Polish	11,204	6.42
ES$_{dobj}$	Romance	18,898	12.81
	Italian	6,375	10.92
ES$_{subj}$	Romance	15,871	7.57
	Italian	5,300	6.98

Table 6: Statistics on the L1 groups and related languages.

L1	Features	Acc	$\mathbf{P}_e$	R_e	F_{1_e}
RU$_{dobj}$	baseline	57.08	51.80	71.58	59.78
	ft$_{En}$	64.20	60.99	55.36	58.04
	+pmi$_{L1}$	65.77	61.06	64.78	62.86
RU$_{subj}$	baseline	56.43	49.52	62.04	54.24
	ft$_{En}$	62.26	55.84	50.02	52.76
	+pmi$_{L1}$	62.78	56.02	54.48	55.21
ES$_{dobj}$	baseline	59.18	51.44	72.31	59.97
	ft$_{En}$	65.14	59.82	53.83	56.66
	+pmi$_{L1}$	66.24	58.92	67.00	62.70
ES$_{subj}$	baseline	58.10	52.95	77.43	62.45
	ft$_{En}$	66.29	61.24	68.45	64.64
	+pmi$_{L1}$	67.00	61.68	70.50	65.78

Table 7: System performance (in %) using L1 and L2 lexico-semantic features, L1 → L1_GROUP.

see for L1 → ALL L1s. Adding the L1 lexico-semantic features has only a minor effect on accuracy and precision, and a more pronounced effect on recall. On the contrary, when we test the system on one particular related L1 (Table 8) we observe the opposite effect: with the exception of ES$_{subj}$ data, precision and accuracy improve, suggesting that the error detection system using L1-induced information identifies errors more precisely.

Overall, the observed gains in performance indicate that L1 semantic models contribute information to lexical choice error detection in L2 for the speakers of typologically related languages. This in turn suggests that there may be less semantic variation within a language group than across different language groups.

7 Discussion and data analysis

The best accuracy achieved in our experiments is 71.19% on ES$_{subj}$ combinations. However, previous research suggests that error detection in lexical choice is a difficult task. For instance, Kochmar and Briscoe (2014) report that the agreement between human annotators on error detection in adjective–noun combinations is 86.50%.

We then qualitatively assessed the performance of our systems by analysing what types of errors

L1	Features	Acc	$\mathbf{P}_e$	R_e	F_{1_e}
RU$_{dobj}$	baseline	55.04	47.68	63.87	53.81
	ft$_{En}$	64.73	59.76	46.05	52.01
	+pmi$_{L1}$	65.15	60.63	45.77	52.16
RU$_{subj}$	baseline	53.30	44.77	61.09	51.29
	ft$_{En}$	61.84	54.63	35.81	43.22
	+pmi$_{L1}$	62.53	57.24	35.11	43.18
ES$_{dobj}$	baseline	55.25	51.67	76.79	61.21
	ft$_{En}$	64.06	62.30	56.01	58.98
	+pmi$_{L1}$	65.21	63.44	58.13	60.66
ES$_{subj}$	baseline	54.34	47.76	68.73	56.23
	ft$_{En}$	62.71	58.80	43.09	49.69
	+pmi$_{L1}$	62.44	58.46	41.71	48.60

Table 8: System performance (in %) using L1 and L2 lexico-semantic features, L1 → REL_L1.

the classifiers reliably detect and what types of errors the classifiers miss across all runs over the samples. Some of the most reliably identified errors in both RU and ES datasets include:

- verbs *offer*, *propose* and *suggest* which are often confused with each other. Correctly identified errors include **offer plan* vs. *suggest plan*, **propose work* vs. *offer work* and **suggest cost* vs. *offer cost*;
- verbs *demonstrate* and *show* where *demonstrate* is often used instead of *show* as in **chart demonstrates*;
- verbs *say* and *tell* particularly well identified with the ES model. Examples include **say idea* instead of *tell idea* and **tell goodbye* instead of *say goodbye*.

These examples represent lexical choice errors when selecting among near-synonyms, and violations of verb subcategorization frames. The error in **find solution* discussed throughout the paper is also reliably identified by the classifier across all runs. It is interesting to note that in the pair of verbs *do* and *make*, which are often confused with each other by both Russian and Spanish L1 speakers, errors involving *make* are identified more reliably than errors involving *do*: for example, **make business* is correctly identified as an error, while **do joke* is missed by the classifier.

Many of the errors missed by the classifier are context-dependent. Some of the most problematic errors involve errors in combinations with verbs like *be* and *become*. Such errors do not result from an L1 lexico-semantic transfer and it is not surprising that the classifiers miss them.

8 Conclusion

We have investigated whether lexico-semantic models from the native language are transferred to the second language, and what effect this transfer has on lexical choice in L2. We focused on two typologically different L1s – Russian and Spanish, and experimentally confirmed the hypothesis that statistical semantic models learned from these L1s significantly improve automatic error detection in L2 data produced by the speakers of the respective L1s. We also investigated whether the semantic models learned from particular L1s are portable to other languages, and in particular to languages that are typologically close to the investigated L1s. Our results demonstrate that L1 models improve the coverage of the error detection system on a range of other L1s.

Acknowledgments

We are grateful to the ACL reviewers for their helpful feedback. Ekaterina Kochmar's research is supported by Cambridge English Language Assessment via the ALTA Institute. Ekaterina Shutova's research is supported by the Leverhulme Trust Early Career Fellowship.

References

Bach E. and Chao W. 2008. *Semantic universals and typology*. In Chris Collins, Morten Christiansen and Shimon Edelman, eds., Language Universals (Oxford: Oxford University Press).

Baldwin T. and Kim S. N. 2010. *Multiword Expressions*. In Handbook of Natural Language Processing, Second Edition, N. Indurkhya and F. J. Damerau (eds.), pp. 267–292.

Ballesteros M., Herrera J., Francisco V., and Gervás P. 2010. *A Feasibility Study on Low Level Techniques for Improving Parsing Accuracy for Spanish Using Maltparser*. In Proceedings of the 6th Hellenic Conference on Artificial Intelligence: Theories, Models and Applications, pp. 39–48.

Baroni M., Bernardini S., Ferraresi A., and Zanchetta E. 2009. *The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora*. Language Resources and Evaluation, 43(3): 209–226.

Briscoe E., Carroll J., and Watson R. 2006. *The Second Release of the RASP System*. In Proceedings of the COLING/ACL-2006 Interactive Presentation Sessions, pp. 59–68.

Burnard L. 2007. *The British National Corpus, version 3 (BNC XML Edition)*. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. `http://www.natcorp.ox.ac.uk/`.

Chang Y.C., Chang J.S., Chen H.J., and Liou H.C. 2012. *An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology*. Computer Assisted Language Learning, 21(3), pp. 283–299.

Dahlmeier D. and Ng H.T. 2011. *Correcting Semantic Collocation Errors with L1-induced Paraphrases*. In Proceedings of the EMNLP-2011, pp. 107–117.

Futagi Y., Deane P., Chodorow M., and Tetreault J. 2009. *A computational approach to detecting collocation errors in the writing of non-native speakers of English*. Computer Assisted Language Learning, 21(4), pp. 353–367.

Joachims T. 1999. *Making Large-Scale SVM Learning Practical. Advances in Kernel Methods – Support Vector Learning*. B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press.

Kochmar E. and Briscoe T. 2014. *Detecting Learner Errors in the Choice of Content Words Using Compositional Distributional Semantics*. In Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers, pp. 1740–1751

Kochmar E. and Briscoe T. 2015. *Using Learner Data to Improve Error Correction in AdjectiveNoun Combinations*. In Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 233-242.

Leacock C., Chodorow M., Gamon M. and Tetreault J. 2014. *Automated Grammatical Error Detection for Language Learners*. Morgan and Claypool Publishers.

Liu A. L.-E., Wible D., and Tsao N.-L. 2009. *Automated suggestions for miscollocations*. In Proceedings of the 4th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 47–50.

Madnani N., Tetreault J., and Chodorow M. 2012. *Exploring Grammatical Error Correction with Not-So-Crummy Machine Translation*. In Proceedings of the 7th Workshop on the Innovative Use of NLP for Building Educational Applications, pp. 44-53.

Mendonca A., Jaquette D., Graff D., and DiPersio D. 2011. *Spanish Gigaword Third Edition*. Linguistic Data Consortium, Philadelphia.

Mikolov T., Sutskever I., Chen K., Corrado G., and Dean J. 2013. *Distributed Representations of Words and Phrases and their Compositionality*. In Proceedings of NIPS.

Mikolov T., Yih W.-T., and Zweig G. 2013. *Linguistic Regularities in Continuous Space Word Representations*. In Proceedings of NAACL HLT.

Miller G. A. 1995. *WordNet: A Lexical Database for English*. Communications of the ACM, 38(11): 39–41.

Mitchell J. and Lapata M. 2008. *Vector-based models of semantic composition*. In Proceedings of ACL, pp. 236–244.

Mitchell J. and Lapata M. 2010. *Composition in distributional models of semantics*. Cognitive Science, 34, pp. 1388–1429.

Nagata, R. and Nakatani, K. 2010. *Evaluating Performance of Grammatical Error Detection to Maximize Learning Effect*. In Proceedings of COLING (Posters), pp. 894-900.

Ng, H.T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., Bryant, C. 2014. *The CoNLL-2014 Shared Task on Grammatical Error Correction*. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task, pp. 1–14.

Nicholls D. 2003. *The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT*. In Proceedings of the Corpus Linguistics conference, pp. 572–581.

Nivre J., Hall J., Nilsson J., Chanev A., Eryigit G., Kübler S., Marinov S., and Marsi E. 2007. *MaltParser: A language-independent system for data-driven dependency parsing*. Natural Language Engineering, 2(13):95–135.

Odlin T. 1989. *Language transfer: Cross-linguistic influence in language learning*. Cambridge University Press.

Östling R. and Knutsson O. 2009. *A corpus-based tool for helping writers with Swedish collocations*. In Proceedings of the Workshop on Extracting and Using Constructions in NLP, NODALIDA, pp. 28–33.

Park T., Lank E., Poupart P., and Terry M. 2008. *Is the sky pure today? AwkChecker: an assistive tool for detecting and correcting collocation errors*. In Proceedings of the 21st annual ACM symposium on User interface software and technology, pp. 121–130.

Rozovskaya A. and Roth D. 2010. *Generating Confusion Sets for Context-Sensitive Error Correction*. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 961-970.

Rozovskaya A. and Roth D. 2011. *Algorithm Selection and Model Adaptation for ESL Correction Tasks*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies – Volume 1, pp. 924–933.

Sharoff S. 2006. *Creating General-Purpose Corpora Using Automated Search Engine Queries*. WaCky! Working papers on the Web as Corpus, Marco Baroni and Silvia Bernardini (ed.).

Sharoff S. and Nivre J. 2011. *The proper place of men and machines in language technology Processing Russian without any linguistic knowledge*. Dialogue 2011, Russian Conference on Computational Linguistics.

Shei C.C. and Pain H. 2000. *An ESL Writer's Collocation Aid*. Computer Assisted Language Learning, 13(2), pp. 167–182.

Vecchi E., Baroni M. and Zamparelli R. 2011. *(Linear) maps of the impossible: Capturing semantic anomalies in distributional space*. In Proceedings of the DISCO Workshop at ACL-2011, pp. 1–9.

Wible H., Kwo C.-H., Tsao N.-L., Liu A., and Lin H.-L. 2003. *Bootstrapping in a language-learning environment*. Journal of Computer Assisted Learning, 19(4), pp. 90–102.