ate translation improvement on corpora of different sizes we conduct experiments with sub-sampled 4K, 8K, and 14K parallel sentences from the training corpora (the smaller the training corpus, the more OOVs it has). Corpora sizes along with statistics of source-side OOV tokens and types are given in tables 1 and 2. Statistics of the held-out dev and test sets used in all translation experiments are given in table 3.

	SW–EN		RO–EN	
	dev	test	dev	test
Sentences	1,552	1,732	2,687	2,265
Tokens	33,446	35,057	24,754	19,659
Types	7,008	7,180	5,141	4,328

Table 3: Dev and test corpora sizes.

In all the MT experiments, we use the `cdec`[9] toolkit (Dyer et al., 2010), and optimize parameters with MERT (Och, 2003). English 4-gram language models with Kneser-Ney smoothing (Kneser and Ney, 1995) are trained using KenLM (Heafield, 2011) on the target side of the parallel training corpora and on the Gigaword corpus (Parker et al., 2009). Results are reported using case-insensitive BLEU with a single reference (Papineni et al., 2002). We train three systems for each MT setup; reported BLEU scores are averaged over systems.

Upper bounds. The goal of our experiments is not only to evaluate the contribution of the OOV dictionaries that we extract when pivoting via borrowing, but also to understand the potential contribution of the lexicon stratification. What is the overall improvement that can be achieved if we correctly translate all OOVs that were borrowed from another language? What is the overall improvement that can be achieved if we correctly translate all OOVs? We answer this question by defining "upper bound" experiments. In the upper bound experiment we word-align all available parallel corpora, including dev and test sets, and extract from the alignments oracle translations of OOV words. Then, we append the extracted OOV dictionaries to the training corpora and re-train SMT setups without OOVs. Translation scores of the resulting system provide an upper bound of an improvement from correctly translating all OOVs. When we append oracle translations of the subset of OOV dictionaries, in particular translations of all OOVs for which the output of the borrowing system is not empty, we obtain an upper bound that can be achieved using our method (if the borrowing system provided perfect outputs). Understanding the upper bounds is relevant not only for our experiments, but for any experiments that involve augmenting translation dictionaries; however, we are not aware of prior work providing similar analysis of upper bounds, and we recommend this as a calibrating procedure for future work on OOV mitigation strategies.

[9] `www.cdec-decoder.org`

Borrowing-augmented setups. As described in §2.2, we integrate translations of OOV loanwords in the translation model. Due to data sparsity, we conjecture that non-OOVs that occur only few times in the training corpus can also lack appropriate translation candidates, i.e., these are target-language OOVs. We therefore run the borrowing system on OOVs and non-OOV words that occur less than 3 times in the training corpus. We list in table 4 sizes of translated lexicons that we integrate in translation tables.

	4K	8K	14K
Loan OOVs in SW–EN	5,050	4,219	3,577
Loan OOVs in RO–EN	347	271	216

Table 4: Sizes of translated lexicons extracted using pivoting via borrowing and integrated in translation models.

Transliteration-augmented setups. In addition to the standard baselines, we evaluate transliteration-augmented setups, where we replace the borrowing model by a transliteration model (Ammar et al., 2012). The model is a linear-chain CRF where we label each source character with a sequence of target characters. The features are label unigrams and bigrams, separately or conjoined with a moving window of source characters. We employ the Swahili–Arabic and Romanian–French transliteration systems that were used as baselines in (Tsvetkov et al., 2015). As in the borrowing system, transliteration outputs are filtered to contain only target language lexicons. We list in table 5 sizes of obtained translated lexicons.

	4K	8K	14K
Translit. OOVs in SW–EN	49	32	22
Translit. OOVs in RO–EN	906	714	578

Table 5: Sizes of translated lexicons extracted using pivoting via transliteration and integrated in translation models.