A Appendix

In order to obtain a reliable part-of-speech (POS) tagging of the `MUSE` test dictionaries efficiently, we used a two-step procedure. First, we ran the Stanford POS tagger (Toutanova et al., 2003) on the English side of each dictionary. We reduced the annotation schema to five categories: nouns (NOUN), proper nouns (PNOUN), verbs (VERB), adjectives and adverbs combined (AD), and others. Next, we asked NLP researchers with the appropriate language background to verify and correct the generated tags, based on both words in a pair. Where one word in the pair is ambiguous with respect to POS, but the other is not, they were told them to use the tag of the latter. If both words were ambiguous, we told them to use the tag they considered more frequent for these words.

We instructed annotators that if a word can be both a proper noun and a common noun, it should be marked as the latter. We told them to mark pairs of identical words as proper nouns, under the assumption that they can be part of a company name or a brand, for example. That is, unless the words in the pair are actual cognates between the source and target language, or they are loanwords. See Table 3 for some examples. Lastly, we asked the annotators to mark pairs as invalid, if the source word is not a valid word in either the source or the target language, or the target word is not a valid translation of the source word. We note that this was a considerable annotation effort if over 40 hours in total. Each annotator had to process over 2000 word pairs: the dictionaries each consist of 1,500 source words, many of which have multiple translations, each processed separately. Annotation was performed in Microsoft Excel.

SRC	TGT	POS	valid	explanation
tea	té	NOUN	✓	actual translation
tea	tea	PNOUN	✓	part of a name, e.g. "Lipton Iced Tea"
rugby	rugby	NOUN	✓	loanword
ugby	ugby	–	✗	not a word in either language

Table 3: Example of annotated gold-standard word pairs from English to Spanish.

B Appendix

The pattern of performance per POS tag is similar for to-EN mappings (see Figure 3), as we saw it for from-EN mapping—proper nouns yield highly variable performance.

Similarly to mappings from-EN, in mappings to-EN (see Figure 4) we see RCSLS outperforming other systems on the clean data for all languages (and by a large margin for most of them), whereas on the original data it appeared inferior to VM-S for DA and HI. Another interesting observation here is that MUSE-U and VM-U occasionally appear inferior to the MUSE-S baseline (for DA and HI, respectively) on the original test data, but on the clean test data all models yield an improvement over the baseline.[11]

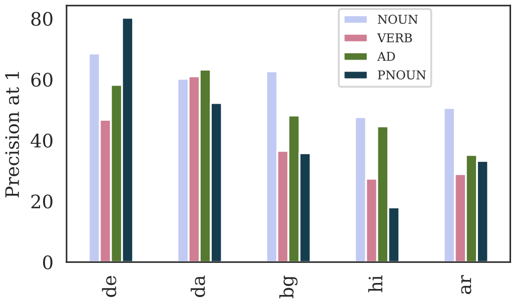

Figure 3: Precision of the RCSLS system, measured per POS tag, on to-EN data.

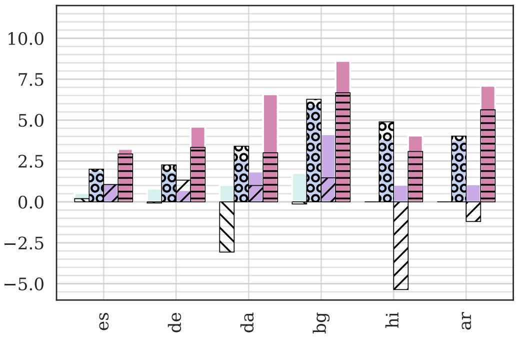

Figure 4: Change in performance on to-EN BDI relative to MUSE-S. Pattern-filled bars show results as estimated on the original data, while colored bars show results as estimated on the cleaned data.

---

[11] That is, excluding MUSE-U evaluated on HI and AR, where all solutions found were degenerate, so they have been excluded.

	es		de		da		bg		hi		ar	
	→ en	en →	→ en	en →	→ en	en →	→ en	en →	→ en	en →	→ en	en →
Source words	1500	1500	1500	1500	1500	1500	1500	1500	1500	1500	1500	1500
	1145	1171	1111	1188	974	1158	1124	1125	963	1104	1212	1080
MUSE-S	83.47	81.66	72.67	73.93	67.07	56.80	56.93	43.93	44.07	33.60	49.93	34.13
	79.56	73.36	66.79	64.47	68.79	55.44	60.63	45.33	46.73	37.68	50.83	34.63
MUSE-U	83.67	82.07	72.60	74.20	64.00	55.40	56.80	39.93	0.00	28.27	0.00	34.60
	80.09	73.78	67.60	64.31	69.82	54.40	62.39	41.51	0.00	34.87	0.00	36.39
VM-S	85.47	81.40	74.93	74.67	70.47	64.60	63.20	48.80	48.96	41.07	53.95	43.53
	81.48	72.50	68.68	65.49	71.46	62.52	66.61	49.78	50.57	45.74	54.62	44.07
VM-U	84.53	82.33	74.00	75.20	68.07	64.87	58.40	44.73	38.71	36.93	48.73	35.73
	80.70	73.53	67.51	65.66	70.64	63.04	64.76	48.44	47.77	44.02	51.90	39.54
RCSLS	86.40	84.46	76.00	79.00	70.07	61.93	63.60	51.73	47.15	38.27	55.56	42.20
	82.79	76.17	71.38	71.97	75.36	62.69	69.24	56.44	50.78	44.57	57.92	45.83

Table 4: Cyan rows correspond to the original test data and white rows to the clean test data. The top rows report the sizes of the dictionaries, measured in terms of source words. For unstable models, e.g. MUSE-U, we train ten models and report results from one random successful model. For a fair comparison of MUSE-U and MUSE-S, we run Procrustes for 5 iterations in both cases, and use the same model selection criterion, mean cosine similarity, in both cases. All systems are evaluated using CSLS for retrieval. * Instead of full annotation for Spanish, we only mark proper nouns and remove them from the test dictionaries to and from English.

C Appendix

	SRC	TGT	RCSLS	VM-S	Description
VM-S ✗, RCSLS ✓	joke	шега лаф виц	шега	шегата	definite form missing from targets
	arbitrators	арбитри	арбитри	арбитрите	definite form missing from targets
	revolt	бунт въстание	бунт	бунта	definite form missing from targets
	remembered	запомнен	запомнен	запомнена	feminine form missing from targets
	hide	скриване	скриване	скриват	*hide* as a verb vs. *hide* as a noun
	bench	пейката пейка	пейка	скамейка	synonym missing from targets
	depot	депо	депо	гара	VM-S predicted ‘station’
	gaelic	келтски	келтски	ирландският	VM-S predicted ‘the irish’
	footage	кадри	кадри	заснети	VM-S predicted ‘shot’
VM-S ✓, RCSLS ✗	egg	яйцето яйца яйце	яйчен	яйце	translation for attributive use of noun missing from targets
	crowned	коронован	коронована	коронован	feminine form missing from targets
	volcanic	вулканична	вулканичен	вулканична	masculine form missing from targets
	penny	пени	паричка	пени	synonym missing from targets
	pound	паунд кг	кило	паунд	RCSLS predicted a non-word
	thursday	четвъртък	петък	четвъртък	RCSLS predicted ‘friday’
	striker	нападател страйкър	защитник	нападател	RCSLS predicted ‘defender’
	pond	езерце	къщичка	езерце	RCSLS predicted ‘cottage’
	flute	флейтата флейта	тромпет	флейта	RCSLS predicted ‘trumpet’
VM-S ✗, RCSLS ✗	circular	кръгло	кръгла	кръгла	feminine form missing from targets
	sailed	отплава	отплавал	отплавал	participle form missing from targets
	grants	субсидии	стипендии	стипендии	synonym missing from targets
	spots	петна	петната	петната	definite form missing from targets
	armies	армии	армиите	армиите	definite form missing from targets
	nose	нос носа носът	врат	задницата	RCSLS predicted ‘neck’, VM-S predicted ‘bottom’
	foods	храни	сладкиши	напитки	RCSLS predicted ‘sweets’, VM-S predicted ’drinks’
	cliff	скала клиф	терас	скалата	RCSLS predicted non-word, definite form missing from targets
	elevated	повишени повишена повишен	понижен	понижен	models predicted ‘reduced’

Table 5: Example translations from EN to BG. In cases where both models predicted forms of the same word, one being more canonical than the other, we underline the canonical form. Truly incorrect translations are marked in grey. Notice the high number of correct translations that are not listed as gold-standard targets.

D Appendix

Table 6 shows an example of an inflectional correspondence map. It signifies that whenever an English word is encountered which is a verb in the infinitive, seven Bulgarian forms would be added to the list of targets, if not in it already. Addition of targets is also conditioned on their presence in the pretrained embeddings vocabulary.

SRC	TGT
V;NINF	V;IMP;2;SG
	V;IMP;2;PL
	V;IND;PRS;1;SG
	V;IND;PRS;1;PL
	V;IND;PRS;2;SG
	V;IND;PRS;2;PL
	V;IND;PRS;3;PL

Table 6: Example of an inflectional correspondence map.

The modifications performed in this manner narrowed the gap in performance between RCSLS and VM-S by only 0.1 percentage points for EN–BG (from 6.7% to 6.4%) and by 1.6 percentage points for EN–DE (from 6.5% to 4.9%). Detailed results can be found in Table 7. Recall that for Bulgarian, we estimated 54% of the gap in performance to stem from false False Positives. If the enrichment procedure was perfect, it should have reduced the gap from 6.6% to less than 3.3%. Unfortunately, due to limited coverage of the inflectional tables and of the pretrained embeddings, only 240 additional word forms were added to the EN–BG dictionary, making for a an almost negligible effect on precision.

	DE	BG
VM-S	65.5	49.8
	67.6	50.3
RCSLS	72.0	56.4
	72.5	56.8
Δ	6.5	6.7
	4.9	6.5

Table 7: Results before (cyan rows) and after (white rows) coverage enrichment for DE and BG