

CharLOTTE: Orthographic Correspondence-Based Data Unification Enhances Cross-Lingual Transfer in Low-Resource NMT

Brendan T. Hatch Peter K. Bickel Eric K. Ringger Stephen D. Richardson

Brigham Young University, USA

{hatch5o6, pbickel, ringger, srichardson}@byu.edu

Abstract

Neural machine translation (NMT) of low-resource languages is often greatly improved through parent-child transfer learning. When high-resource parent and low-resource child languages are closely related, the system benefits from high vocabulary overlap and shared representations. However, systematic surface-level differences in orthography hide many of the rich connections between related languages, even when written in the same script, and prevent knowledge transfer. We present CharLOTTE: a completely automatic and language-agnostic pipeline that overcomes this hurdle without the need of manually crafted rules or external linguistic resources. CharLOTTE collects a dataset of cognates by running word alignment on parallel sentences and filtering the resulting word pairs based on normalized edit distance. The pipeline learns systematic orthographic correspondences from the cognates and applies them to more closely unify the representation of a high-resource language to that of a related low-resource language, thus increasing vocabulary overlap and enhancing knowledge transfer that improves downstream NMT. We evaluate CharLOTTE on four low-resource translation pairs, including one creole language pair, and report significant improvement in translation quality of up to 5.53 BLEU.

1 Introduction

In transfer-learning scenarios of neural machine translation (NMT), Nguyen and Chiang (2017) proposed that a single vocabulary represent parent and child languages so that the child model can leverage shared token embeddings that were learned in the parent model. Vocabulary overlap has also shown to help in certain other NLP tasks (Limisiewicz et al., 2023; Patil et al., 2022), with Patil et al. (2022) emphasizing that the overlap matters in scenarios where languages are related, especially when one is at low resource. This work

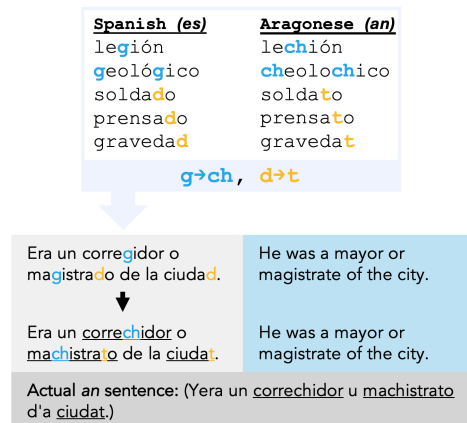


Figure 1: $g \rightarrow ch$ and $d \rightarrow t$ correspondence patterns are highlighted in the cognate pairs. Below, the first row shows an $es \rightarrow en$ sentence pair, and the second row shows the reshaped $es \rightarrow en$ pair. For reference, the an version of the source sentence is provided with exact word matches with the reshaped es underlined.

increases vocabulary overlap by leveraging orthographic correspondences to reshape a high-resource language to exhibit orthographic patterns of a related low-resource language, enhancing knowledge transfer. We obviate the need for manually crafted transliteration rules and external language-specific resources, presenting instead a completely automated pipeline that requires only parallel data.

Key to our method is the automatic detection and analysis of cognates: words from different languages that share an ancestral origin. Cognates are similar in meaning, and usually bear resemblance in form *with systematic differences*. When analyzing cognate pairs between two related languages, sound correspondences are found; that is, sounds in certain environments in one language correspond systematically to different sounds in the other language. Many of these correspondences are reflected in orthography. Figure 1 shows Spanish (es) to Aragonese (an) cognates exhibiting correspondences $g \rightarrow ch$ and $d \rightarrow t$. These differences are

not random, but occur in predictable environments.

We present CharLOTTE (**Character-Level Orthographic Transfer for Translation Enhancement**)¹, a language-agnostic pipeline that applies learned orthographic correspondences to data unification, reshaping a high resource language to exhibit orthographic patterns of a related low-resource language, thus enhancing knowledge transfer in low-resource NMT. We evaluate this method in four low-resource scenarios that yield improved translation quality. While we adopt a parent-child transfer learning strategy, our data unification method is not limited to any one knowledge transfer paradigm.

2 Related Work

2.1 Transfer Learning in Low-Resource NMT

Many works develop transfer learning methods for low-resource NMT. Zoph et al. (2016) initialize models with parent models trained on a high-resource language pair and Maimaiti et al. (2019) do similarly, but in a multi-round fashion with multiple high-resource parent languages. Roy et al. (2024) initialize seq2seq encoders with pretrained multilingual models while employing complementary knowledge distillation to correct training imbalance, and Gu et al. (2018) enable multiple languages to share word and sentence-level representations. Shurtz et al. (2025) embed transliterations to align related languages of different scripts in the embedding space, and other works leverage pivot languages (Cheng et al., 2017; Kim et al., 2019).

2.2 Data Unification

Unifying textual representations is key to enabling cross-lingual transfer. Nguyen and Chiang (2017) and Shurtz et al. (2025) transliterate languages of different scripts to a common script. Maimaiti et al. (2019) introduce a manually crafted unified transliteration that converts Arabic-script languages into a common character representation. They map individual characters across script variants to reduce superficial character-set differences, but do not capture the deeper, context-dependent orthographic correspondences (*e.g.*, $g \rightarrow ch$ in particular environments) that arise between related same-script languages. Shurtz et al. (2025) create a cognate-aware transliteration and Currey et al. (2016) apply com-

mon correspondence patterns; but again, both use manually crafted character transformation rules. These prior efforts represent a progression from cross-script unification (Nguyen and Chiang, 2017) to manual phonetic character mapping within a script family (Maimaiti et al., 2019), to manually crafted cognate-aware transliteration (Shurtz et al., 2025). Other works automate data unification, such as Pourdamghani and Knight (2017, 2019) who use weighted finite state transducers trained on monolingual sentences to apply character mappings to a high-resource language, converting it into a related low-resource language, and thus augmenting low-resource parallel data for statistical machine translation (SMT). CharLOTTE advances this line of work by applying neural models to data unification and evaluating its benefit in downstream neural machine translation (NMT).

2.3 Cognate Detection

The field of computational historical linguistics has developed many methods for automatic cognate detection. Early methods include Kondrak (2001) and Kondrak (2009), which approach the problem with measures for semantic and phonetic similarity. LexStat (List, 2012) and InfoMap (List et al., 2017) operate on multilingual word lists. Other methods include a shingling scheme (A, 2018); edit-distance approaches enhanced with phonological features, character embeddings, and cognate labels (McCoy and Frank, 2018); bipartite networks (Rama and List, 2019); and pointwise mutual information-based strategies (Rama et al., 2017; Dellert, 2018). Several studies also employ SVMs and other classical machine learning algorithms (Jäger et al., 2017; Ciobanu and Dinu, 2014; Labat and Lefever, 2019; Dinu et al., 2023). More recently, neural-based methods for cognate detection are used (Rama, 2016; Soisalon-Soininen and Granroth-Wilding, 2019; Lefever et al., 2020; Kanojia et al., 2019, 2020; Goswami et al., 2023; Akavarapu and Bhat-tacharya, 2024; Kanojia et al., 2021).

We use an inexpensive method to extract lists of cognate pairs from parallel sentences. It is also designed to work on orthography rather than phonetic transcriptions. Methods like these appear in Grönroos et al. (2018); Malmasi and Dras (2015); Labat and Lefever (2019); Lefever et al. (2020); Kanojia et al. (2019), which often identify word alignments from parallel data and filter them down to word pairs within some edit distance threshold of each other, yielding a list of candidate cognate

¹Code to train our models and replicate our work is made available at <https://github.com/byu-matrix-lab/CharLOTTE1.0>.

pairs that are similar in form and meaning.

2.4 Sound Correspondence Detection and Cognate Prediction

There are many methods for automatic sound correspondence detection (Kondrak, 2002, 2009; List, 2019; Fourier and Sagot, 2020a; Blum and List, 2023), and the related task of cognate prediction (Fourier and Sagot, 2020a; Beinborn et al., 2013; Fourier et al., 2021; Fourier and Sagot, 2022; Zundui et al., 2025). These cognate prediction methods frame the task as a machine translation problem, which requires the application of sound correspondence patterns learned from the cognate pairs.

2.5 Cognate-Informed Machine Translation

Leveraging cognates in machine translation is not new. Mann and Yarowsky (2001) apply transduction models of cognate pairs to induce translation lexicons. Al-Onaizan et al. (1999) use cognates as a training constraint and as additional parallel training data in SMT, and Kanojia et al. (2020) use them as additional parallel training data in NMT. We enable cognate-informed machine translation by learning more granular orthographic correspondences to unify high and low-resource languages.

3 The CharLOTTE Pipeline

We refer to two kinds of language pairs: (1) cognate language pairs, which are a pair of related languages from which cognates are extracted to learn data unification patterns, and (2) translation language pairs. Cognate language pairs will be notated as $L1/L2$, as in *es/an*. Translation language pairs will be notated as $L1 \rightarrow L2$, as in *an \rightarrow en*.

We use the term *parent language* to mean a high-resource language for training a *parent* model, and *child language* to mean a related low-resource language for training a *child* model (i.e., the parent is **not** necessarily an etymological ancestor). We consider scenarios where we have a parent language PL , a related child language CL , and a target language TL , where the final objective is to translate the low-resource pair $CL \rightarrow TL$ leveraging knowledge transfer from the high-resource pair $PL \rightarrow TL$ enhanced by applying PL/CL orthographic correspondences. The data resources required to use CharLOTTE include $PL \rightarrow TL$, $CL \rightarrow TL$, and $PL \rightarrow CL$ parallel sentences.

This section describes the CharLOTTE pipeline, (see Figure 2), which is summarized as follows:

1. Extract PL/CL cognates from $PL \rightarrow CL$ parallel sentences.
2. Use cognates to train a PL/CL orthographic correspondence (OC) model.
3. Apply the PL/CL OC model to the PL side of the $PL \rightarrow TL$ parallel data, yielding $PL' \rightarrow TL$ data.
4. Train a $CL \rightarrow TL$ NMT model, leveraging knowledge transfer from $PL' \rightarrow TL$.

3.1 Extracting Cognates from Parallel Data

First, CharLOTTE creates a dataset of unique cognate word pairs between PL and CL to train a PL/CL OC model. While detecting true cognates is complicated, the pipeline employs the following heuristic that simply considers similarity in meaning and form: given $PL \rightarrow CL$ sentences,

1. Make word alignments from the parallel sentences with Fast Align (Dyer et al., 2013)² using grow-diag-final-and symmetrization. This controls for similarity in meaning.
2. Filter word pairs down to those with a normalized Levenshtein distance (NLD) less than or equal to some θ . This controls for similarity in form. Pairs of identical words are kept.

3.2 OC Models

The resulting cognate dataset is used to train an OC model that applies PL/CL orthographic correspondences. This is framed as a character-level machine translation task with the PL words as source sequences and the CL words as targets, teaching the model to apply context-dependent character mappings. Because identical word pairs (NLD = 0.0) are kept in the training data, models also learn when transformations are not needed.

To train OC models, we adopt CopperMT³ (Fourier et al., 2021), a framework for training cognate prediction models using NMT or SMT architectures. While the authors trained models on sequences of phones, we train models on orthography, which therefore learn *orthographic correspondences* rather than true *sound correspondences*.

Once trained, the PL/CL OC model is run on each word of the PL side of the $PL \rightarrow TL$ NMT training, validation, and test sets. While the task is

²https://github.com/clab/fast_align.git

³<https://github.com/clefouier/CopperMT>

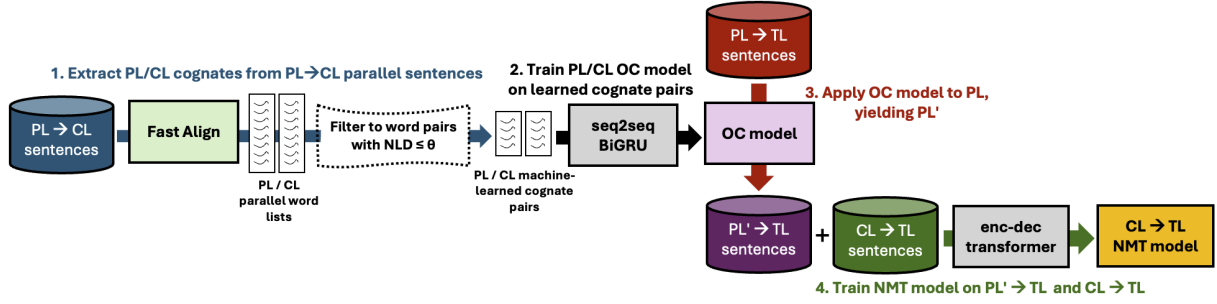


Figure 2: The CharLOTTE pipeline: (1) Cognate pairs are mined by running $PL \rightarrow CL$ sentences through Fast Align and filtering the resulting word pairs to those with a normalized Levenshtein distance (NLD) $\leq \theta$. (2) These cognates are used to train an OC model. (3) The OC model then transforms each PL word in the $PL \rightarrow TL$ sentences. This results in $PL' \rightarrow TL$ sentences for (4) enhanced knowledge transfer to the $CL \rightarrow TL$ NMT task.

called cognate prediction, a better way to frame this is the rewriting of PL with orthographic patterns of CL . Not all PL words actually have cognates in CL , but the model is still applied. This yields a new language PL' , and sets of $PL' \rightarrow TL$ sentences.

3.3 NMT With OC-Based Data Unification

CharLOTTE then trains a transformer NMT model with parent-child transfer learning, pretraining a parent model on $PL' \rightarrow TL$ and fine-tuning its checkpoint on $CL \rightarrow TL$ to create the child model. The encoder and decoder are initialized with the same vocabulary built from CL , PL' , and TL .

We also consider the reverse translation scenarios from a source language SL into CL . These use the same sets of parallel sentences described, but flip the translation direction, making TL the SL . We pretrain NMT models on $SL \rightarrow PL'$ and fine-tune on $SL \rightarrow CL$. While the translation direction is flipped, the cognate direction is the same, so the same PL' sentences are reused.

4 Experiments

We notate experimental scenarios as $PL/CL \rightarrow TL$, as in $es/an \rightarrow en$, where the objective is to translate $CL \rightarrow TL$ ($an \rightarrow en$) with transfer learning from $PL' \rightarrow TL$ ($es' \rightarrow en$) where PL' (es') is PL (es) reshaped by PL/CL (es/an) OC mappings. We notate reverse scenarios as $SL \rightarrow PL/CL$, as in $en \rightarrow es/an$ where the objective is to translate $SL \rightarrow CL$ ($en \rightarrow an$) with transfer learning from $SL \rightarrow PL'$ ($en \rightarrow es'$).

We perform experiments with four scenarios and their respective reverse scenarios: $es/an \rightarrow en$, $fr/mfe \rightarrow en$, $fr/oc \rightarrow en$, and $uz/kaa \rightarrow en$, where es = Spanish, an = Aragonese, en = English, fr = French, mfe = Mauritian Creole, oc = Occitan, uz = Uzbek, and kaa = Karakalpak.

Table 1 summarizes the parallel data and Appendix A describes data sources and preparation.

Each PL/CL is chosen for shared ancestry and linguistic features. es , an , fr and oc are Romance languages (Hammarström et al., 2026), a family with SVO word order (Dryer, 2013) and a characteristic fusional morphology. As a creole, mfe 's primary lexifier is fr (Baker and Kriegel, 2013), and while having simpler morphology than fr , it has the same SVO word order (Baker and Kriegel, 2013). uz and kaa are Turkic languages (Hammarström et al., 2026) with SOV word order (Dryer, 2013) and similar agglutinative morphology.

4.1 Training OC Models

Lists of cognate pairs are created with the method in Section 3.1. Following Labat and Lefever (2019) and Lefever et al. (2020), we set $\theta = 0.5$. The lists are divided into OC training and validation sets on a 90/10% split, with a max validation size of 1,100. Cognate pairs with duplicate source words are removed from the validation set. Then, training pairs containing either source or target words that exist in the validation split are removed. Table 12 shows the final OC dataset sizes.

The authors of CopperMT (Fourrier et al., 2021) report that RNNs (seq2seq BiGRU) perform better on bilingual cognate prediction than transformers, with SMT sometimes best in low-resource scenarios. Using the CopperMT framework, we train OC models with a seq2seq Bi-GRU architecture implemented with fairseq (Ott et al., 2019). Table 15 reports fixed hyperparameters, and Table 14 reports hyperparameters tuned in a grid search, with the best model configurations reported in Table 13. We compared performance to the CopperMT SMT system, implemented with Moses (Koehn et al., 2007) and GIZA++ (Och and Ney, 2003). We found that

PL/CL→TL	nPL/CL	nPL→TL			nCL→TL		
		Train	Val.	Test	Train	Val.	Test
<i>es/an</i> → <i>en</i>	28,594	9,999,993	997	1,012	20,089	997	1,012
<i>fr/mfe</i> → <i>en</i>	13,319	9,999,989	997	1,012	19,140	472	811
<i>fr/oc</i> → <i>en</i>	8,767,173	9,999,989	997	1,012	1,536,211	997	1,012
<i>uz/kaa</i> → <i>en</i>	85,845	9,999,997	997	1,012	50,895	2,000	1,012

Table 1: **nPL/CL** is the # of *PL/CL* parallel sentences for cognate mining. **nPL→TL** and **nCL→TL** are the # of parent and child parallel sentences for NMT training, validation and testing.

PL/CL	nGT	PL words	PL' words
<i>es/an</i>	406	77.47	83.44
<i>fr/mfe</i>	315	31.93	77.01
<i>fr/oc</i>	365	55.57	60.88

Table 2: charBLEU of the original *PL* words versus the *PL'* words (OC output) scored against the *CL* cognates from the ground truth test set, which has a size of **nGT**.

BiGRU was superior to SMT for all scenarios based on validation BLEU (Papineni et al., 2002; Post, 2018) calculated over characters (charBLEU)⁴.

While OC model performance is ultimately tested in downstream NMT, where possible, we conducted an intermediary stage performance test by collecting ground truth sets of cognate pairs (see Appendix A.4 for details), running the OC models on the *PL* words, and calculating charBLEU against the *CL* cognates. As a baseline, we calculate charBLEU of the original *PL* words. Table 2 reports results for *es/an*, *fr/mfe*, and *fr/oc* models, where it is observed that the OC model greatly succeeds in making *PL* words more like *CL* words, especially when going from *fr* to *mfe*.

4.2 Applying OC Models

The OC model is then applied to the *PL* → *TL* sentences. Table 16 shows examples of *PL* words and the generated *PL'* versions. We measure the change from *PL* to *PL'* with a series of metrics applied to the NMT *PL* → *TL* training set.

Table 3 reports the raw difference with three metrics: S-NLD is the average character-level NLD between each *PL'* sentence and the corresponding *PL* sentence. %WΔ is the percentage of *PL'* words that are different from their corresponding *PL* words. For *PL'* words that are different, W-NLD is the average NLD between them and their corresponding *PL* words.

⁴Used the SacreBLEU "char" tokenizer.

PL/CL→TL	S-NLD	%WΔ	W-NLD
<i>es/an</i> → <i>en</i>	0.05	20.22	0.21
<i>fr/mfe</i> → <i>en</i>	0.29	76.51	0.41
<i>fr/oc</i> → <i>en</i>	0.08	31.89	0.28
<i>uz/kaa</i> → <i>en</i>	0.25	73.33	0.35

Table 3: Differences between the created *PL'* and the original *PL* sentences from the *PL* → *TL* NMT training set as a result of applying the *PL/CL* OC model.

```
$ _ g e n e r a l i z a d a m e n t e $
$ c h e n e r a l i z a d a m e n t _ $
n-gram mappings: g→h, _g→ch, _ge→che, ...
```

Figure 3: Levenshtein alignment between a *PL* word (top) and the *PL'* word. \$ = word boundary; _ = gap.

In addition, we capture the meaningful character mappings between *PL* and *PL'*, reflecting the OC model’s learned orthographic correspondences. We start with the Levenshtein alignment between each unique *PL* word and its corresponding *PL'* word, as in Figure 3. For each character-level *n*-gram for $1 \leq n \leq 5$, we count the number of instances it is mapped into each target *n*-gram. Applying this to each *PL*, *PL'* word pair creates a frequency distribution of the mappings for each *PL* *n*-gram.

We use Shannon entropy to measure the uncertainty of how each *PL* *n*-gram transforms, meaning a *PL* *n*-gram with low entropy has a consistent mapping in *PL'*. The Shannon entropy *H* for a character *n*-gram is defined as:

$$H(X) = - \sum_{i=1}^N P(x_i) \log_2 P(x_i) \quad (1)$$

where $P(x_i)$ is the probability of the *n*-gram being mapped into the *i*-th possible mapping.

Measuring entropy at different values of *n* shows what contexts give consistent mappings. Table 4 shows mapping distributions and entropy of three

PL n -gram	Frequency Distribution	H
g	g:556, h:176, l:4	0.84
_g	ch:151, ll:2	0.1
_ge	che:74	0.0

Table 4: Illustrative example of frequency distributions of PL n -gram mappings to PL' , and the entropy (H).

PL n -grams, each a sub-string of the subsequent n -gram. As more context is considered, we discover lower-entropy mappings. Notably, $_ge \rightarrow che$ has a more consistent context than $_g \rightarrow ch$. Additionally, the highest entropy n -gram g most often maps to itself, which is not a transformation.

Ergo, we developed a process to tease apart the n -gram mappings that meaningfully reflect regular transformations from those that do not, which works as follows: We remove the gap characters $_$ resulting from the Levenshtein alignment and merge the mapping frequency distributions accordingly (e.g., merge $_g$ and g mappings). Then, we filter out mappings where the PL n -gram most often maps to itself, occurs infrequently, has an entropy over a mathematically determined threshold, or whose mapping is a complete composition of other low-entropy mappings (see Appendix B for details). Random samples of meaningful n -gram mappings are given in Tables 21, 22, 23, 24.

Total counts of the meaningful n -gram mapping types that result, as well how frequently they occur in the unique list of PL , PL' word pairs, and how frequently they are applied to the entire $PL \rightarrow TL$ NMT training set, are given in Table 5. We also present their density, measured by normalizing the respective raw frequencies over the total number of characters in the PL sides of the word list (word density) and $PL \rightarrow TL$ sentences (applied density). We expect that high densities of meaningful mappings applied to the $PL \rightarrow TL$ data will correspond with improved $CL \rightarrow TL$ translation, and low densities with little to no improvement.

4.3 NMT Models

Following Nguyen and Chiang (2017), for each scenario, a single tokenizer is trained on the union of the parent language, child language, and target language, i.e., PL' , CL , and TL . We use SentencePiece (Kudo and Richardson, 2018)⁵ BPE as our tokenization algorithm.

⁵<https://github.com/google/sentencepiece>

With these vocabularies, we train NMT models with the BartForConditionalGeneration (Lewis et al., 2020)⁶ transformer architecture with modified configurations based on settings of transformer-base (Vaswani et al., 2017).

For each scenario, a parent NMT model is trained with the $PL' \rightarrow TL$ training and validation sets. Once converged, its best checkpoint is saved and training is resumed on the $CL \rightarrow TL$ training and validation sets. The final child model is evaluated on the $CL \rightarrow TL$ test set.

For evaluation, we create a transfer-learning baseline and a simple baseline. The transfer-learning baseline uses parent-child transfer learning like the proposed method, but with PL instead of PL' . Likewise, its tokenizer is trained on PL , CL , and TL (not PL'). The simple baseline is trained without transfer learning on only the $CL \rightarrow TL$ training and validation data, though it uses the same tokenizer as the transfer-learning baseline.

Tokenizers, architectural configurations, and training hyperparameters for proposed and baseline NMT models are detailed in Appendix C.

4.4 Evaluation

4.4.1 Vocabulary Overlap

We hypothesize that CharLOTTE’s improvements stem in part from greater vocabulary overlap between PL' and CL than between PL and CL , which occurs because PL' exhibits more of the orthographic patterns of CL . We examine this hypothesis with two complementary metrics, though acknowledging there are likely other contributing factors as well. We measure vocabulary overlap with Jensen-Shannon divergence (JSD) (Lin, 1991), following Limisiewicz et al. (2023), and with Jaccard similarity (formal definitions in Appendix D).

Low JSD scores indicate similar token frequency distributions whereas high Jaccard similarity scores indicate high overlap of unique token types. We take the token sets and token distributions of the PL , PL' , and CL sentences from the $PL \rightarrow TL$, $PL' \rightarrow TL$, and $CL \rightarrow TL$ NMT training sets, and report the Jaccard similarity and JSD between PL and CL , and between PL' and CL in Table 6. CharLOTTE decreases JSD for three out of four scenarios, only slightly increasing JSD between fr and oc . However, Jaccard similarity always

⁶https://huggingface.co/docs/transformers/en/model_doc/bart#transformers.BartForConditionalGeneration – We clarify that we are using only the architecture and *not* the pretrained weights.

PL/CL $\rightarrow$ TL	Type		Word		Applied	
<i>es/an</i> $\rightarrow$ <i>en</i>	0.5K	1.8K	9.5K / 0.001	0.2M / 0.023	1.1M / 0.001	27M / 0.027
<i>fr/mfe</i> $\rightarrow$ <i>en</i>	7.5K	20.7K	173.6K / 0.019	1.7M / 0.177	13.8M / 0.012	221M / 0.198
<i>fr/oc</i> $\rightarrow$ <i>en</i>	0.9K	2.8K	18.1K / 0.002	0.2M / 0.021	1.1M / 0.001	20M / 0.018
<i>uz/kaa</i> $\rightarrow$ <i>en</i>	4.7K	14.5K	128.7K / 0.008	1.3M / 0.082	2.5M / 0.004	52M / 0.082

Table 5: Counts of meaningful n -gram mappings with 0 entropy (left) and of *all* meaningful n -gram mappings (right) as a result of applying the *PL/CL* OC model. **Type** is the # of unique mappings. **Word** is the # of times these mappings occur in the unique list of *PL*, *PL'* word pairs, presented as count / density. **Applied** is the # of times these mappings are applied to the entire *PL* $\rightarrow$ *TL* NMT training set, presented as count / density.

Data	JSD $\downarrow$	Jaccard $\uparrow$
<i>es</i> – <i>an</i>	0.242	0.608
<i>es'</i> – <i>an</i>	0.209 (- 0.033)	0.678 (+ 0.07)
<i>fr</i> – <i>mfe</i>	0.465	0.218
<i>fr'</i> – <i>mfe</i>	0.341 (- 0.124)	0.418 (+ 0.2)
<i>fr</i> – <i>oc</i>	0.243	0.706
<i>fr'</i> – <i>oc</i>	0.261 (+ 0.018)	0.725 (+ 0.019)
<i>uz</i> – <i>kaa</i>	0.397	0.467
<i>uz'</i> – <i>kaa</i>	0.199 (- 0.198)	0.719 (+ 0.252)

Table 6: Jensen-Shannon divergence and Jaccard similarity to measure vocabulary overlap between *PL* and *CL* compared to *PL'* and *CL* (deltas in parentheses).

increases, suggesting that CharLOTTE primarily increases overlap of the token sets rather than fully aligning token frequency distributions.

4.4.2 Translation Quality

We measure translation quality with BLEU and chrF (Popović, 2015), calculated with SacreBLEU (Post, 2018)⁷ with paired approximate randomization (Riezler and Maxwell, 2005) for statistical significance. Table 7 has scores for translating to and from *en*. Table 8 reports BLEURT-20 (Sellam et al., 2020; Pu et al., 2021) for $\rightarrow$ *en* directions. CharLOTTE is the strongest system for all pairs, with the greatest benefit for *mfe* $\leftrightarrow$ *en* and *uz* $\leftrightarrow$ *en* translation, consistently shown with large increases in BLEU, chrF, and BLEURT-20. The simple baseline is always weakest by great margins.

Parent-child transfer learning is of great benefit, but is enhanced with CharLOTTE. We believe this is because many of the rich relationships between related languages occur at finer-grained levels than sub-word tokens, which obscure these connections. CharLOTTE bypasses this problem, unifying par-

ent and child languages and increasing vocabulary overlap. This is especially true with knowledge transfer from *fr* to *mfe* and *uz* to *kaa*. Table 3 shows that these two OC models make the greatest changes to the parent languages, changing well over 70% of words for each (%W Δ). The degree to which the *fr/mfe* and *uz/kaa* OC models change those words is also greatest, with W-NLD of 0.41 and 0.35 (versus 0.21 and 0.28). Furthermore, Table 5 shows that the applied density of meaningful n -gram mappings applied by the *fr/mfe* and *uz/kaa* models (0.198 and 0.082) is much greater than that applied by the *es/an* and *fr/oc* models (0.027 and 0.018). Consequently, the greatest increases in vocabulary overlap occur between *fr* and *mfe*, and *uz* and *kaa*. The parent languages have to be radically reshaped to optimize the number of shared tokens with the child languages. This makes sense when considering, for instance, that *mfe*, despite being a French-based creole, follows very different spelling conventions. The benefit of the large *fr/mfe* and *uz/kaa* transformations is noted in the greatest payoffs in translation.

Additionally, while CharLOTTE yields the greatest gains for the three lowest resource pairs *mfe* $\leftrightarrow$ *en*, *uz* $\leftrightarrow$ *en*, and *an* $\leftrightarrow$ *en*, it also improves translation for *oc* $\rightarrow$ *en*, a pair with 1.5M sentences. This improvement occurs despite an increase in JSD between *fr* and *oc* after applying CharLOTTE, suggesting that unique token overlap (reflected in higher Jaccard similarity) and the resulting embedding sharing may matter more than alignment of token frequency distributions for transfer learning in this setting, though the inconsistency across metrics indicates that vocabulary overlap alone may not fully account for CharLOTTE’s gains. Other potential factors, such as improved sub-word segmentation and the corresponding shared embeddings, merit investigation in future work.

That said, there is still a correlation between the

⁷<https://github.com/mjpost/sacrebleu>

PL/CL→TL	nP/C	nC→T	Simple	Transfer	CharLOTTE	Δ
<i>es/an</i> → <i>en</i>	48K	20K	1.43/20.05	18.15/47.84	**†20.55/**†50.72	+ 2.4/2.88
<i>fr/mfe</i> → <i>en</i>	6.1K	19K	4.31/24.81	15.11/36.43	**18.38/**42.01	+ 3.27/5.58
<i>fr/oc</i> → <i>en</i>	1.2M	1.5M	11.76/36.08	26.68/53.37	*28.22/**55.13	+ 1.54/1.76
<i>uz/kaa</i> → <i>en</i>	108K	51K	1.99/24.49	15.79/44.36	**21.32/**49.27	+ 5.53/4.91
<i>en</i> → <i>es/an</i>	48K	20K	0.88/19.53	11.14/43.04	*12.44/**45.44	+ 1.3/2.4
<i>en</i> → <i>fr/mfe</i>	6.1K	19K	5.97/22.88	15.44/37.51	**19.94/**43.94	+ 4.5/6.43
<i>en</i> → <i>fr/oc</i>	1.2M	1.5M	5.57/28.64	16.98/44.63	17.52/†45.43	+ 0.54/0.8
<i>en</i> → <i>uz/kaa</i>	108K	51K	1.01/20.90	8.39/40.85	**12.01/**46.90	+ 3.62/6.05
<i>esx/an</i> → <i>en</i>	111K	20K	1.43/20.05	18.15/47.84	†† 19.17/**49.42	+ 1.02/1.58

Table 7: BLEU/chrF scores for simple and transfer baselines, and for CharLOTTE. **nP/C** is the # of OC training cognate pairs. **nC→T** is the # of *CL* → *TL* parallel training sentences (all scenarios had 10M *PL* → *TL*). *esx/an* → *en* is the ablation from Section 4.4.3. Δ is the difference between CharLOTTE and plain transfer. Scores significant over the transfer baseline are indicated with ** (p < 0.0001), * (p < 0.0005), and †† (p < 0.005), and † (p < 0.05). ‡ indicates that *es/an* → *en* CharLOTTE is significant over (p < 0.0001) *esx/an* → *en* CharLOTTE.

PL/CL→TL	S	T	Ch
<i>es/an</i> → <i>en</i>	0.2956	0.5811	0.6223
<i>fr/mfe</i> → <i>en</i>	0.3866	0.5141	0.5883
<i>fr/oc</i> → <i>en</i>	0.4493	0.5764	0.5926
<i>uz/kaa</i> → <i>en</i>	0.3561	0.5667	0.6288
<i>esx/an</i> → <i>en</i>	0.2956	0.5811	0.6042

Table 8: BLEURT-20 for simple (S), transfer (T), and CharLOTTE (Ch) → *en* systems. All Ch scores are significant over T scores (p = 0.0001), as is *es/an* → *en* Ch over *esx/an* → *en* Ch.

change in vocabulary overlap due to CharLOTTE and NMT performance, with large changes in JSD and Jaccard similarity (between *fr* and *mfe*, *uz* and *kaa*) correlating with large NMT gains, moderate changes (between *es* and *an*) correlating with moderate NMT gains, and small changes (between *fr* and *oc*, even negatively by JSD) with modest NMT gains. This also correlates with the fact that applying the *fr/oc* OC model results in the lowest density of applied meaningful *n*-gram mappings (0.018). This occurs despite the *fr/oc* → *en* scenario having greater S-NLD, %WΔ, and W-NLD metrics than *es/an* → *en*, meaning that the *fr* data is changed more than the *es* data is, even though fewer meaningful *n*-gram mappings are applied. This suggests that the *fr/oc* model may be applying more noise than the other OC models, perhaps due to dirtier training data.

To confirm that CharLOTTE improves NMT, we conducted a human rank evaluation between

the CharLOTTE and transfer-learning baseline systems. For each of the *an* → *en* and *kaa* → *en* language pairs, four native English speakers were presented with the reference translations from 200 random samples of the respective test set along with the CharLOTTE and transfer-learning baseline translations. The evaluators compared the system translations to the reference translation and indicated the better translation with a score of 1, and the other with 2, with ties allowed. Figure 5 shows the instructions and interface. The sentence scores were then averaged, yielding the *rank* metric, a number bound between 1 and 2 where lower values indicate overall better translations. We report the average of the four evaluators' *rank* metrics in Table 10, along with the average number of times each system scored better than the other system (wins), and the number of times they received the same score (ties). Individual evaluator scores are reported in Table 20, where it is observed that all evaluators consistently had strong preference for the CharLOTTE system.

To measure inter-annotator agreement, we converted the sentence-level annotations to categories "CharLOTTE" or "transfer" according to which system received a win, or "tied" if both systems received the same score, and calculated Fleiss' Kappa (Fleiss, 1971)⁸ over these three categories. We also report Krippendorff's Alpha (Krippendorff, 2013)⁹ with an ordinal distance metric with value do-

⁸Seabold and Perktold (2010), link to docs [here](#).

⁹Castro (2017), use `level_of_measurement='ordinal'` and `value_domain=[0, 1, 2]`.

PL/CL→TL	κ	α	Unanimous	Soft	Hard w/Tie	Hard
<i>es/an</i> → <i>en</i>	0.33	0.48	59 (29.5%)	112 (56.0%)	21 (10.5%)	8 (4.0%)
<i>uz/kaa</i> → <i>en</i>	0.36	0.47	83 (41.5%)	58 (29.0%)	31 (15.5%)	28 (14.0%)

Table 9: Inter-annotator agreement for the rank evaluation of the CharLOTTE and transfer systems. κ = Fleiss’ Kappa; α = Krippendorff’s Alpha; **Unanimous** = the # of sentences (and % of total sentences) where every annotator gave the same score, **Soft** = where ties and preferences for the same system are given, **Hard w/Tie** = where ties and preferences for both systems are given, **Hard** = where preferences for both systems are given without ties.

Model	Rank ↓	Wins ↑	Ties
<i>es/an</i> → <i>en</i>			77.25
CharLOTTE	1.38	88.25	
Transfer	1.65	34.5	
<i>uz/kaa</i> → <i>en</i>			35.0
CharLOTTE	1.36	116.5	
Transfer	1.70	48.5	

Table 10: Human *rank* scores averaged across four evaluators for the CharLOTTE and transfer systems, as well as the number of wins and ties.

main of $[0, 1, 2]$, where 0="transfer", 1="tied", and 2="CharLOTTE", thereby applying smaller penalty to soft disagreements with the "tied" category (e.g., "CharLOTTE" vs. "tied") than to hard disagreements with an opposing system category (e.g., "CharLOTTE" vs. "transfer"). Table 9 reports these metrics along with the number of hard and soft disagreements. For both pairs, Fleiss’ Kappa indicates "fair" ($0.21 \leq \kappa \leq 0.40$) agreement (Lan-dis and Koch, 1977), which is typical for similar NMT evaluations (Toral and Way, 2018).

4.4.3 θ -Ablation

To determine how important the threshold θ is when mining cognates from the parallel word lists, we conducted an ablation with the *es/an* → *en* scenario by setting $\theta = 1.0$, effectively removing the filter. We refer to this ablated scenario as *esx/an* → *en*. The resulting OC training size is 111K, compared to 48K when $\theta = 0.5$ (see Table 12). We conducted the OC model hyperparameter search from Section 4.1 and report the best parameters in Table 13. While there is a great disparity in the OC validation charBLEU scores between the *es/an* and *esx/an* OC models (86.41 vs. 42.60), we have to acknowledge that the *esx/an* validation set was sampled from both a distinct and noisier set of word pairs and the only way to compare these two scenarios is in downstream NMT (see

Tables 7 and 8). The *esx/an* → *en* CharLOTTE model still outperforms its transfer-learning baseline, but at a smaller margin of 1.02 BLEU rather than 2.4 BLEU. This leaves a significant gap of 1.38 BLEU between the *es/an* → *en* CharLOTTE NMT model and the *esx/an* → *en* CharLOTTE NMT model. This suggests that while CharLOTTE can still learn useful correspondence patterns from unfiltered lists of words, filtering improves its ability to do so.

5 Conclusion

This work presents the language-agnostic method CharLOTTE to enhance cross-lingual transfer in low-resource NMT. CharLOTTE mines cognates from parallel data and from them learns orthographic correspondences to unify the representations of related languages and increase vocabulary overlap. This yields measurably significant benefits for low-resource translation, especially of *mfe* and *kaa* due to radical transformation of the parent languages which enhanced knowledge transfer.

Future work may apply this method to more languages from diverse language families and scripts to further test its generalizability. Options for mining cognate pairs that do not rely on pre-existing sets of *PL* → *CL* parallel sentences should also be explored to make the method more widely applicable. Future work may also compare the meaningful *PL*, *PL'* *n*-gram mappings to real-life documented sound correspondences between the respective *PL* and *CL*. Additionally, noting that the *fr/oc* model likely applies unwanted noise, characterization of the induced noise in a manner complementary to the meaningful *n*-gram mappings should be developed. Finally, future work could explore measuring meaningful and irregular *n*-gram mappings between a *PL* and *CL* prior to running the pipeline as a way to predict if CharLOTTE-enhanced transfer is likely to occur.

Limitations

This work evaluates CharLOTTE on four language scenarios. We acknowledge the limits of these data points and look forward to future work applying this method to a wider range of languages of various families and scripts to test generalizability. We suspect that if few regular character mappings exist between a parent and child language, or if a child language has a vocabulary with significant numbers of loan words from various languages, then OC models may struggle to learn transformations that are consistently useful. We leave it to future work to explore this further.

In addition, we only analyzed the increase in vocabulary overlap—the *amount* of shared tokens—as one explanation for CharLOTTE’s gains, but there are likely other factors at play. We hypothesize that OC models unify the *PL* and *CL* textual representations such that the resulting tokenizer’s sub-word boundaries are more optimally aligned with corresponding morphological features of the two languages. That is, we suspect that more morphemes that are essentially the same but surface slightly differently in each language are assigned the same token representation. However, this has not yet been investigated.

Finally, while applying OC models proves to be of benefit in downstream NMT, we suspect they also introduce some degree of noise, which is especially possible if the *PL* $\rightarrow$ *CL* parallel sentences are dirty. If they contain misaligned sentences or incorrect languages, this can lead to a dirty set of extracted cognate pairs which will make it difficult for the OC model to learn regular orthographic correspondences. Analyzing the meaningful *n*-gram metrics against the S-NLD, %W Δ , W-NLD metrics across the different scenarios suggested to us this may be at play in the *fr/oc* $\rightarrow$ *en* scenario, but we have not yet thoroughly examined the possibility. Future work should therefore characterize noisy *n*-gram mappings in a manner complementary to the meaningful *n*-gram mappings.

Ethical Considerations

AI was used to assist in coding, and to suggest wordings and apply minor edits to the manuscript.

References

Pranav A. 2018. [Alignment analysis of sequential segmentation of lexicons to improve automatic cognate](#)

[detection](#). In *Proceedings of ACL 2018, Student Research Workshop*, pages 134–140, Melbourne, Australia. Association for Computational Linguistics.

V.S.D.S.Mahesh Akavarapu and Arnab Bhattacharya. 2024. [Automated cognate detection as a supervised link prediction task with cognate transformer](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 965–975, St. Julian’s, Malta. Association for Computational Linguistics.

Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah A Smith, and David Yarowsky. 1999. Statistical machine translation. In *Final Report, JHU Workshop 1999*.

Philip Baker and Sibylle Kriegel. 2013. [Mauritian creole structure dataset](#). In Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors, *Atlas of Pidgin and Creole Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Khuyagbaatar Batsuren, Gabor Bella, and Fausto Giunchiglia. 2019. [CogNet: A large-scale cognate database](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3136–3145, Florence, Italy. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2022. A large and evolving cognate database. *Language Resources and Evaluation*, 56(1):165–189.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2013. [Cognate production using character-based machine translation](#). In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 883–891, Nagoya, Japan. Asian Federation of Natural Language Processing.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O’Reilly Media Inc.

Frederic Blum and Johann-Mattis List. 2023. [Trimming phonetic alignments improves the inference of sound correspondence patterns from multilingual wordlists](#). In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 52–64, Dubrovnik, Croatia. Association for Computational Linguistics.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff’s alpha agreement measure. <https://github.com/pln-fing-udelar/fast-krippendorff>.

Yong Cheng, Yang Liu, Qian Yang, Maosong Sun, and Wei Xu. 2017. [Joint training for pivot-based neural machine translation](#). *Preprint*, arXiv:1611.04928.

- Alina Maria Ciobanu and Liviu P. Dinu. 2014. [Automatic detection of cognates using orthographic alignment](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 99–105, Baltimore, Maryland. Association for Computational Linguistics.
- Anna Currey, Alina Karakanta, and Jon Dehdari. 2016. [Using related languages to enhance statistical language models](#). In *Proceedings of the NAACL Student Research Workshop*, pages 116–123, San Diego, California. Association for Computational Linguistics.
- Johannes Dellert. 2018. [Combining information-weighted sequence alignment and sound correspondence models for improved cognate detection](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3123–3133, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Liviu P. Dinu, Ioan-Bogdan Iordache, and Ana Sabina Uban. 2023. [CoToHiLi at SIGTYP 2023: Ensemble models for cognate and derivative words detection](#). In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 137–142, Dubrovnik, Croatia. Association for Computational Linguistics.
- Matthew S. Dryer. 2013. [Order of subject, object and verb \(v2020.4\)](#). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.
- Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. [A simple, fast, and effective reparameterization of IBM model 2](#). In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. [Beyond english-centric multilingual machine translation](#). Preprint, arXiv:2010.11125.
- Joseph L. Fleiss. 1971. [Measuring nominal scale agreement among many raters](#). *Psychological Bulletin*, 76(5):378–382.
- Clémentine Fourrier, Rachel Bawden, and Benoît Sagot. 2021. [Can cognate prediction be modelled as a low-resource machine translation task?](#) In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 847–861, Online. Association for Computational Linguistics.
- Clémentine Fourrier and Benoît Sagot. 2020a. [Comparing statistical and neural models for learning sound correspondences](#). In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 79–83, Marseille, France. European Language Resources Association (ELRA).
- Clémentine Fourrier and Benoît Sagot. 2020b. [Methodological aspects of developing and managing an etymological lexical resource: Introducing EtymDB-2.0](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3207–3216, Marseille, France. European Language Resources Association.
- Clémentine Fourrier and Benoît Sagot. 2022. [Probing multilingual cognate prediction models](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3786–3801, Dublin, Ireland. Association for Computational Linguistics.
- Koustava Goswami, Priya Rani, Theodorus Franssen, and John McCrae. 2023. [Weakly-supervised deep cognate detection framework for low-resourced languages using morphological knowledge of closely-related languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 531–541, Singapore. Association for Computational Linguistics.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10.
- Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2018. [Cognate-aware morphological segmentation for multilingual neural translation](#). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 386–393, Belgium, Brussels. Association for Computational Linguistics.
- Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. [Universal neural machine translation for extremely low resource languages](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.
- Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2026. [Glottolog 5.3](#). Available online at <http://glottolog.org>, Accessed on 2026-08-25.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spacy: Industrial-strength natural language processing in python](#).
- Gerhard Jäger, Johann-Mattis List, and Pavel Sofroniev. 2017. [Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists](#). In *Proceedings of the 15th Conference of the European Chapter of*

- the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1205–1216, Valencia, Spain. Association for Computational Linguistics.
- Diptesh Kanojia, Raj Dabre, Shubham Dewangan, Pushpak Bhattacharyya, Gholamreza Haffari, and Malhar Kulkarni. 2020. [Harnessing cross-lingual features to improve cognate detection for low-resource languages](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1384–1395, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Diptesh Kanojia, Kevin Patel, Malhar Kulkarni, Pushpak Bhattacharyya, and Gholamreza Haffari. 2019. [Utilizing wordnets for cognate detection among Indian languages](#). In *Proceedings of the 10th Global Wordnet Conference*, pages 404–412, Wrocław, Poland. Global Wordnet Association.
- Diptesh Kanojia, Prashant Sharma, Sayali Ghodkar, Pushpak Bhattacharyya, Gholamreza Haffari, and Malhar Kulkarni. 2021. [Cognition-aware cognate detection](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3281–3292, Online. Association for Computational Linguistics.
- Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. [Pivot-based transfer learning for neural machine translation between non-English languages](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 866–876, Hong Kong, China. Association for Computational Linguistics.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. [Moses: Open source toolkit for statistical machine translation](#). In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.
- Grzegorz Kondrak. 2001. [Identifying cognates by phonetic and semantic similarity](#). In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.
- Grzegorz Kondrak. 2002. [Determining recurrent sound correspondences by inducing translation models](#). In *COLING 2002: The 19th International Conference on Computational Linguistics*.
- Grzegorz Kondrak. 2009. [Identification of cognates and recurrent sound correspondences in word lists](#). In *Traitement Automatique des Langues, Volume 50, Numéro 2 : Langues anciennes [Ancient Languages]*, pages 201–235, France. ATALA (Association pour le Traitement Automatique des Langues).
- Klaus Krippendorff. 2013. *Content Analysis: An Introduction to its Methodology*, 3rd edition. Sage, Thousand Oaks, CA.
- Taku Kudo and John Richardson. 2018. [SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.
- Sofie Labat and Els Lefever. 2019. [A classification-based approach to cognate detection combining orthographic and semantic similarity information](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 602–610, Varna, Bulgaria. INCOMA Ltd.
- J. Richard Landis and Gary G. Koch. 1977. [The measurement of observer agreement for categorical data](#). *Biometrics*, 33(1):159–174.
- Els Lefever, Sofie Labat, and Pranaydeep Singh. 2020. [Identifying cognates in English-Dutch and French-Dutch by means of orthographic information and cross-lingual word embeddings](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4096–4101, Marseille, France. European Language Resources Association.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Tomasz Limisiewicz, Jiří Balhar, and David Mareček. 2023. [Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5661–5681, Toronto, Canada. Association for Computational Linguistics.
- J. Lin. 1991. [Divergence measures based on the shannon entropy](#). *IEEE Transactions on Information Theory*, 37(1):145–151.
- Pierre Lison and Jörg Tiedemann. 2016. [OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).
- Johann-Mattis List. 2012. [LexStat: Automatic detection of cognates in multilingual wordlists](#). In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 117–125, Avignon, France. Association for Computational Linguistics.

- Johann-Mattis List. 2019. [Automatic inference of sound correspondence patterns across multiple languages](#). *Computational Linguistics*, 45(1):137–161.
- Johann-Mattis List, Simon J. Greenhill, and Russell D. Gray. 2017. [The potential of automatic word comparison for historical linguistics](#). *PLOS ONE*, 12(1):1–18.
- Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. [Effective approaches to attention-based neural machine translation](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.
- Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, and Maosong Sun. 2019. [Multi-round transfer learning for low-resource nmt using multiple high-resource languages](#). *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18(4).
- Shervin Malmasi and Mark Dras. 2015. [Cognate identification using machine translation](#). In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 138–141, Parramatta, Australia.
- Mukhammadsaid Mamasaidov and Abror Shopulatov. 2024. [Open language data initiative: Advancing low-resource machine translation for Karakalpak](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 606–613, Miami, Florida, USA. Association for Computational Linguistics.
- Gideon S. Mann and David Yarowsky. 2001. [Multipath translation lexicon induction via bridge languages](#). In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.
- R. Thomas McCoy and Robert Frank. 2018. Phonologically informed edit distance algorithms for word alignment with low-resource languages. *Society for Computation in Linguistics*, 1(1).
- Toan Q. Nguyen and David Chiang. 2017. [Transfer learning across low-resource, related languages for neural machine translation](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. [Scaling neural machine translation to 200 languages](#). *Nature*, 630(8018):841–846.
- Franz Josef Och and Hermann Ney. 2003. [A systematic comparison of various statistical alignment models](#). *Computational Linguistics*, 29(1):19–51.
- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. [fairseq: A fast, extensible toolkit for sequence modeling](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Vaidehi Patil, Partha Talukdar, and Sunita Sarawagi. 2022. [Overlap-based vocabulary generation improves cross-lingual transfer among related languages](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–233, Dublin, Ireland. Association for Computational Linguistics.
- Juan Antonio Perez-Ortiz, Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, Aaron Galiano Jimenez, Antoni Oliver, Claudi Aventín-Boya, Alejandro Pardos, Cristina Valdés, Jusép Loís Sans Socasau, and Juan Pablo Martínez. 2024. Expanding the flores+ multilingual benchmark with translations for Aragonese, Aranese, Asturian, and Valencian. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, USA. Association for Computational Linguistics.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Nima Pourdamghani and Kevin Knight. 2017. [Deciphering related languages](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2513–2518, Copenhagen, Denmark. Association for Computational Linguistics.
- Nima Pourdamghani and Kevin Knight. 2019. [Neighbors helping the poor: improving low-resource machine translation using related languages](#). *Machine Translation*, 33(3):239–258.
- Amy Pu, Hyung Won Chung, Ankur P Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for mt. In *Proceedings of EMNLP*.

- Taraka Rama. 2016. [Siamese convolutional networks for cognate identification](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1018–1027, Osaka, Japan. The COLING 2016 Organizing Committee.
- Taraka Rama and Johann-Mattis List. 2019. [An automated framework for fast cognate detection and Bayesian phylogenetic inference in computational historical linguistics](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6225–6235, Florence, Italy. Association for Computational Linguistics.
- Taraka Rama, Johannes Wahle, Pavel Sofroniev, and Gerhard Jäger. 2017. Fast and unsupervised methods for multilingual cognate clustering. *arXiv preprint arXiv:1702.04938*.
- Stefan Riezler and John T. Maxwell. 2005. [On some pitfalls in automatic evaluation and significance testing for MT](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, Michigan. Association for Computational Linguistics.
- Nathaniel Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome Etori, Vijay Murari Tiyyala, Olanrewaju Samuel, Matthew Stutzman, Bismarck Odoom, Sanjeev Khudanpur, Stephen Richardson, and Kenton Murray. 2024. [Kreyòl-MT: Building MT for Latin American, Caribbean and colonial African creole languages](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3083–3110, Mexico City, Mexico. Association for Computational Linguistics.
- Aniruddha Roy, Pretam Ray, Ayush Maheshwari, Sudeshna Sarkar, and Pawan Goyal. 2024. [Enhancing low-resource NMT with a multilingual encoder and knowledge distillation: A case study](#). In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 64–73, Bangkok, Thailand. Association for Computational Linguistics.
- Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. [Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia](#). *Preprint*, arXiv:1907.05791.
- Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2020. [Ccmatrix: Mining billions of high-quality parallel sentences on the web](#). *Preprint*, arXiv:1911.04944.
- Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.
- Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.
- Ammon Shurtz, Christian Richardson, and Stephen D. Richardson. 2025. [When scripts diverge: Strengthening low-resource neural machine translation through phonetic cross-lingual transfer](#). In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 336–346, Suzhuo, China. Association for Computational Linguistics.
- Bernard W. Silverman. 1986. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall/CRC, London.
- Eliel Soisalon-Soininen and Mark Granroth-Wilding. 2019. [Cross-family similarity learning for cognate identification in low-resource languages](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1121–1130, Varna, Bulgaria. INCOMA Ltd.
- Jörg Tiedemann. 2016. [Finding alternative translations in a large corpus of movie subtitles](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 3518–3522, Portorož, Slovenia. European Language Resources Association (ELRA).
- Antonio Toral and Andy Way. 2018. [What Level of Quality Can Neural Machine Translation Attain on Literary Text?](#), pages 263–287. Machine Translation: Technologies and Applications. Springer International Publishing AG.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, and 16 others. 2020. [SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python](#). *Nature Methods*, 17(3):261–272.
- Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. [Transfer learning for low-resource neural machine translation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.
- Tsolmon Zundui, Khuyagbaatar Batsuren, Tsendsuren Munkhdalai, and Amarsanaa Ganbold. 2025. [Cognate production using character-based neural machine translation without segmentation](#). *IEEE Access*, 13:34824–34830.

A Data

A.1 Sources of Parallel Sentences

Each of the four $PL/CL \rightarrow TL$ scenarios evaluated in this work rely on sets of $PL \rightarrow CL$, $PL \rightarrow TL$, and $CL \rightarrow TL$ parallel sentences, for a total of 12 sets of parallel data. In this section, we will discuss the data sources for each of these scenarios. The respective reverse $SL \rightarrow PL/CL$ scenarios discussed in Section 3.3 use the same sets, just flipping the direction of the $PL \rightarrow TL$ and $CL \rightarrow TL$ sentences.

Table 11 shows the sources of the parallel sentences used in each $PL/CL \rightarrow TL$ scenario, whose references and details are listed here:

- **OpenSubtitles** (Lison and Tiedemann, 2016; Tiedemann, 2016) – <https://opus.nlpl.eu/OpenSubtitles/corpus/version/OpenSubtitles>, created from data extracted from <http://www.opensubtitles.org/>
- **WikiMatrix** (Schwenk et al., 2019) – <https://opus.nlpl.eu/WikiMatrix/corpus/version/WikiMatrix>
- **FLORES+** (NLLB Team et al., 2024; Goyal et al., 2022; Perez-Ortiz et al., 2024; Mamasaidov and Shopulatov, 2024) – https://huggingface.co/datasets/openlanguagedata/flores_plus, Aragonese expansion: https://huggingface.co/datasets/openlanguagedata/flores_plus/blob/main/dataset_cards/arg_Latn.md, Karakalpak extension: https://huggingface.co/datasets/openlanguagedata/flores_plus/blob/main/dataset_cards/kaa_Latn.md
- **CCMatrix** (Schwenk et al., 2020; Fan et al., 2020) – <https://opus.nlpl.eu/CCMatrix/corpus/version/CCMatrix>
- **Kreyol-MT** (Robinson et al., 2024) – <https://huggingface.co/jhu-clsp/kreyol-mt>
- **NLLB** (Schwenk et al., 2020; Fan et al., 2020) – <https://opus.nlpl.eu/NLLB/corpus/version/NLLB>
- **OLDI (Karakalpak)** (Mamasaidov and Shopulatov, 2024) – <https://huggingface.co/datasets/tahrirchi/dilmash>

A.2 Parallel Data Preparation and Cleaning

For our parent language pairs, $es \rightarrow en$, $fr \rightarrow en$ (used for both $fr/mfe \rightarrow en$ and $fr/oc \rightarrow en$ scenarios) and $uz \rightarrow en$, we retrieved 10M parallel sentences from the original dataset. For all other sets, we retrieved the full set. For each language pair, we then removed any sentence pairs in the training set whose source segment or target segment also occurred in the validation and test data. Then, with exception of $fr \rightarrow mfe$ and $mfe \rightarrow en$, we cleaned each parallel training set with a pipeline that follows the guidelines of the GILT Leaders Forum’s Best Practices in Translation Memory Management¹⁰, and performs the following steps:

1. Remove pairs containing empty source or target segments.
2. Remove pairs when the source segment exactly or nearly matches the target segment.
3. Remove duplicate source-target pairs.
4. Remove pairs with segments containing mostly non-alphabetic characters.
5. Remove pairs with segments containing abnormally long sequences of characters without spaces, including segments that are only URLs.
6. Remove pairs containing segments with unbalanced brackets.
7. Remove pairs containing fewer than 3 words in the English source segment.
8. Remove pairs with segments containing a higher number of characters than 5 standard deviations above the mean for that language (sentences that are too long).
9. Remove pairs in which the ratio of the lengths of the source and target segments exceeds a certain cutoff.
10. Normalize escaped Unicode characters.
11. Validate and normalize character encodings for each language.
12. Normalize whitespace
13. Shorten sequences of excessively repeated punctuation.
14. Normalize quotation marks.
15. Normalize HTML entities.
16. Remove all markup tags.

¹⁰<https://github.com/GILT-Forum/TM-Mgmt-Best-Practices/blob/master/best-practices.md>

Pair	Cognate Extraction / NMT Train	NMT Val./Test
<i>es/an</i> → <i>en</i>		
<i>es</i> → <i>an</i>	OpenSubtitles, WikiMatrix	-
<i>es</i> → <i>en</i>	CCMatrix	FLORES+ (dev/devtest)
<i>an</i> → <i>en</i>	OpenSubtitles, WikiMatrix	FLORES+ (dev/devtest)
<i>fr/mfe</i> → <i>en</i>		
<i>fr</i> → <i>mfe</i>	Kreyol-MT (train)	-
<i>fr</i> → <i>en</i>	CCMatrix	FLORES+ (dev/devtest)
<i>mfe</i> → <i>en</i>	Kreyol-MT (train)	Kreyol-MT (validation/test)
<i>fr/oc</i> → <i>en</i>		
<i>fr</i> → <i>oc</i>	NLLB	-
<i>fr</i> → <i>en</i>	CCMatrix	FLORES+ (dev/devtest)
<i>oc</i> → <i>en</i>	NLLB	FLORES+ (dev/devtest)
<i>uz/kaa</i> → <i>en</i>		
<i>uz</i> → <i>kaa</i>	OLDI	-
<i>uz</i> → <i>en</i>	NLLB	FLORES+ (dev/devtest, <i>Northern Uzbek</i>)
<i>kaa</i> → <i>en</i>	OLDI	OLDI / FLORES+ (devtest)

Table 11: Sources for parallel sentences for each $PL/CL \rightarrow TL$ scenario. Note that for $kaa \rightarrow en$, there is no FLORES+ dev set for validation, and we therefore removed a 2,000-sentence split from the OLDI data to serve as the validation set.

PL/CL	Train	Val.
<i>es/an</i>	47,655	1,093
<i>fr/mfe</i>	6,148	702
<i>fr/oc</i>	1,178,753	1,090
<i>uz/kaa</i>	107,711	1,096
<i>esx/an</i>	110,637	1,021

Table 12: For each PL/CL OC model, the number of cognate pairs in the training and validation sets created by the CharLOTTE pipeline. *esx/an* refers to the θ -ablation in Section 4.4.3

We opted not to clean the $fr \rightarrow mfe$ and $mfe \rightarrow en$ parallel sentences because they contained a great many single-word parallel sentences that would have been removed by the cleaning pipeline. We acknowledge that best practice removes short sentence pairs, but given that these were small datasets already, we were concerned about the data loss that would have resulted.

Final dataset sizes are reported in Table 1.

A.3 OC Data Preprocessing

Prior to aligning words for OC training, we pre-tokenize the PL and CL sentences by applying NLTK (Bird et al., 2009) and spaCy (Honni-

bal et al., 2020) word tokenizers. Prior to OC inference, the same tokenizers and additional cleaning steps, like removing trailing and leading punctuation, are applied to the PL sentences to create the list of words fed to the OC model to be transformed into PL' . When constructing a PL' sentence from the transformed words, we do so in a way that preserves the original PL sentence’s spacing, formatting, and punctuation. We used the following word tokenizers for each language:

- *en*: NLTK word_tokenize, "english"
- *fr*: NLTK word_tokenize, "french"
- *mfe*: NLTK word_tokenize, "french"
- *es*: spaCy es_core_news_sm
- *an*: spaCy es_core_news_sm
- *oc*: spaCy es_core_news_sm
- *uz*: spaCy xx_sent_ud_sm
- *kaa*: spaCy xx_sent_ud_sm

A.4 Ground Truth Cognate Test Data

To create ground truth cognate test sets to evaluate the OC models, we extracted cognate pairs from

PL/CL	Layers	Emb. Dim.	Hid. Dim.	Batch Size	charBLEU
<i>es/an</i>	2	128	512	256	86.41
<i>fr/mfe</i>	2	32	256	64	86.90
<i>fr/oc</i>	6	32	512	512	65.97
<i>uz/kaa</i>	2	64	256	512	83.19
<i>esx/an</i>	6	128	256	512	42.60

Table 13: The best seq2seq BiGRU configurations for each *PL/CL* OC model from hyperparameter grid search. **Layers** refers to the number of both encoder and decoder layers. **charBLEU** refers to the character-level BLEU on the validation set. *esx/an* refers to the θ -ablation in Section 4.4.3.

Parameter	Values
layers	2, 4, 6
embed. dim.	16, 32, 64, 128
hidden dim.	16, 32, 64, 128, 256, 512
batch size	16, 64, 256, 512

Table 14: The grid search space for the seq2seq Bi-GRU OC models. Each OC model trained on one NVIDIA A100. Training time varied from a few minutes to up to one day depending on the dataset and hyperparameters.

Parameter	Value
Attention	Luong-dot
Learning rate	0.001
Dropout	0.2

Table 15: Fixed hyperparameters for the seq2seq Bi-GRU OC models. "Luong-dot" refers to the dot attention from Luong et al. (2015).

the CogNet (Batsuren et al., 2019, 2022)¹¹ and EtymDB (Fourrier and Sagot, 2020b)¹² datasets.

To extract cognates from EtymDB, we used a modified version of the CopperMT extractor_script_cognates.py¹³ script, along with modified language_info.py¹⁴. For Aragonese ancestors, we used ine-pro, itc-pro, itc-ola, la, lat-vul. For Mauritian Creole ancestors, we used ine-pro, itc-pro, itc-ola, la, lat-vul, ofr, frm, fr. We will include our versions of these scripts in our published code.

We then sanitized the cognate sets by removing

¹¹<https://github.com/kbatsuren/CogNet>

¹²<https://github.com/clefourrier/EtymDB>

¹³https://github.com/clefourrier/CopperMT/blob/master/pipeline/data/extractor_script_cognates.py

¹⁴https://github.com/clefourrier/CopperMT/blob/master/pipeline/data/management/from_etymdb/utils/language_info.py

any cognate pairs where either the source or target word occurs in the OC training or validation data, and removed duplicate cognate pairs. This resulted in sets of 406 *es/an*, 315 *fr/mfe*, and 365 *fr/oc* cognate pairs.

B Meaningful n -Gram Mappings

We explain our process for extracting meaningful n -gram mappings from the list of raw n -gram mapping frequency distributions.

B.1 Insertions and Deletions

First, we must properly account for gap characters before finalizing the n -gram mapping frequency distribution. The gap character `_` is an artifact of the alignment process representing insertions and deletions and should be removed. This means that the counts for the n -gram `_g` should be combined with the counts for the n -gram `g` in a way that does not double count mappings. For example, if `_g→ch: 40`, `g→h: 45`, then the combined tally should be `g→ch: 40`, `h: 5` instead of `g→ch: 40`, `h: 45`.

We can ensure that there is no double counting of such mappings with the following process: When considering an n -gram mapping, if the source n -gram (from *PL*) is preceded or followed by a gap character, then we consider the mapping incomplete because an insertion is occurring in the immediately greater context. Therefore its count is not included at the current context size of n . For example, consider the bigram mapping `ge→he` in the *PL*, *PL'* alignment `$_gen$, $schen$` (where `$` = word boundary). The *PL* bigram `ge` comes after a gap character, so the correct mapping should be counted as `ge→che`, which is processed at the trigram level `_ge→che`. After removing the gap character, we correctly tally counts for `ge→che` without double counts of the `ge→he` mapping. Similarly, if there is a gap character preceding or following an output

PL/CL	PL Word	PL' Word
<i>es/an</i>	responsabilidades	responsabilidatz
	posteriores	posteriors
	acostarse	acostar-se
	pareció	pareixió
	derecha	dreita
<i>fr/mfe</i>	l'épaisseur	lepeser
	française	fransez
	stérilité	sterilite
	copiant	kopian
	étions	esion
<i>fr/oc</i>	organisationnels	organizacionals
	administrée	administrada
	dissimulé	dissimulat
	moderne	moderna
	élever	elever
<i>uz/kaa</i>	kesishdi	kesisti
	yengilsa	jeñilse
	yelpig'ich	jelpigish
	tabiatlarini	tábiyatların
	trayektoriasida	traektoriyasında

Table 16: Examples of *PL* words and the corresponding *PL'* words (*PL/CL* OC model predictions.)

n -gram (from *PL'*), a deletion is occurring in the immediately greater context and the n -gram should not be counted at the current context size. For example, consider the trigram mapping $\text{ent} \rightarrow \text{ent}$ in the alignment $\text{\$mente\$}$, $\text{\$ment_\$}$. The correct mapping is actually $\text{ente} \rightarrow \text{ent}$, which is processed at the 4-gram level $\text{ente} \rightarrow \text{ent_}$.

B.2 Filtering

We filter the data down to the most meaningful mappings. We apply four filters:

First, meaningful mappings must pass a non-identity criterion, meaning we filter out mappings where the *PL* n -gram maps to itself in half or more of its output n -grams. This allows us to see the actual transformations that the OC model makes.

Second, we remove mappings where the input n -grams falls below a frequency threshold, which is set to 10 in our analysis. This filters out rare occurrences, ensuring that what we consider as a meaningful mapping follows an established pattern.

Third, we calculate entropy and remove mappings whose entropy is too high to be considered meaningful. The zero-entropy mappings reflect correspondences that the model has applied with-

out exception, thus these are all kept. We can classify the rest of the mappings as one of two types: consistent correspondences with a small amount of noise (noisy correspondences), such as $\text{ad\$} \rightarrow \text{at\$}$:72, $\text{ad\$}$:1 (entropy = 0.1), or highly variable mappings without the context for a clear determination, such as $\text{did} \rightarrow \text{dit}$:10, did :9 (entropy = 1.0). When qualitatively reviewing the mappings, it appeared to us that the entropy threshold between the noisy correspondences and the variable mappings was somewhere midway between 0 and 1. We plotted the entropy distributions (Figure 4) of the n -gram mappings for each language scenario and consistently noted clear valleys around the 0.5 mark. We associate these valleys with the distinction between the noisy correspondences and variable mappings. We formalize this visual threshold by fitting the entropy distribution with Kernel Density Estimation (KDE) using SciPy's (Virtanen et al., 2020) `scipy.stats.gaussian_kde`¹⁵ module, scaling the Silverman bandwidth (Silverman, 1986) by a factor of 0.5 to preserve the major peaks and valleys of the distribution. The KDE was eval-

¹⁵https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.gaussian_kde.html

uated on a 1000-point grid spanning the observed entropy range, and local minima were identified via `scipy.signal.find_peaks`¹⁶ on the negated KDE, requiring a prominence of at least 5% of the maximum density. We exclude zero-entropy mappings from the KDE, as their mass would artificially bias the distribution toward zero. We take the entropy value at the leftmost prominent local minimum as our threshold, which separates what appear to be the first two modes. Assuming that the first mode represents noisy correspondences, we filter out mappings whose entropy is greater than this threshold. The resulting KDE-based entropy thresholds are visualized in Figure 4. Interestingly, the threshold for all scenarios is roughly 0.55.

To assess sensitivity to bandwidth scaling values and prominence percentage requirements, we recomputed the threshold for each language pair across Silverman bandwidth scales from 0.2 to 0.8 and prominence percentages from 1% to 10% of the max distribution density. With prominence percentages and bandwidth scales at the highest end of these ranges for *es/an*, *fr/oc*, and *uz/kaa*, KDE is unable to find any prominent local minima, which highlights the need to use a small enough scaling factor to recognize the multimodal distribution and prevent oversmoothing. For all combinations in this range that do find prominent minima, the resulting thresholds varied by at most 0.068 within any pair (*es/an*: 0.533–0.551; *fr/mfe*: 0.525–0.593; *fr/oc*: 0.539–0.565; *uz/kaa*: 0.515–0.563). Thus, beyond using a small enough bandwidth to recognize the multimodal distribution, there is little sensitivity to the specific values of the KDE and prominence parameters.

Fourth, we filter out mappings that are completely composed of other low-entropy mappings. Consider the data in Table 17. The mapping of *átic*→*atic* is not meaningful because it is already accounted for by the combination of the two smaller mappings. So, if an *n*-gram mapping with entropy below our threshold can be fully explained by a composition of smaller *n*-gram mappings with entropy below our threshold, then we filter the larger *n*-gram mapping out. Note that for these explanatory mappings, we only filter by the entropy threshold and frequency and not by the non-identity criterion. Otherwise, *átic*→*atic* would be consid-

PL	Frequency Distribution	<i>H</i>
<i>át</i>	<i>at</i> :1481, <i>át</i> :82, ...	0.41
<i>ic</i>	<i>ic</i> :50426, <i>__</i> :162, ...	0.09
<i>átic</i>	<i>atic</i> :1228, <i>_tic</i> :9, ...	0.25

Table 17: Example of a redundant *PL* *n*-gram mapping to *PL'*. *átic*→*atic* is completely accounted for by the other two mappings. *H* is entropy.

ered meaningful because *ic*→*ic* has been filtered out from the first filtering step. In summary, we filter out *n*-gram mappings that do not meet all of the following 4 criteria:

1. **Non-Identity:** The *PL* *n*-gram must not map to itself in half or more of its output *n*-grams.
2. **Frequency:** The *PL* *n*-gram must occur at least 10 times.
3. **Entropy:** The mapping must be below the entropy threshold.
4. **Redundancy:** The mapping must not be fully explained by a combination of other meaningful mappings.

Random samplings of the final meaningful *PL*, *PL'* *n*-gram mappings for each language scenario are given in Tables 21, 22, 23, and 24.

C NMT Model Details

C.1 Tokenizers

Each CharLOTTE NMT model tokenizer is trained on the union of the parent language, child language, and target language, i.e. *PL'*, *CL*, and *TL*. We sample sentences from *PL' → CL*, *PL' → TL*, and *CL → TL* datasets. Sentences for each language are up-/down-sampled so that the total tokenizer training data size is 500K sentences with the *PL' : CL : TL* ratio being 1:1:2. These sentences are then used to train a SentencePiece (Kudo and Richardson, 2018)¹⁷ BPE tokenizer with character coverage of 1.0 and vocabulary size of 32K.

The tokenizer for each transfer-learning baseline is made with the same process, but using *PL* sentences instead of *PL'* in the tokenizer training data. The simple baselines use the same tokenizers as the respective transfer-learning baselines.

¹⁶https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.find_peaks.html

¹⁷<https://github.com/google/sentencepiece>

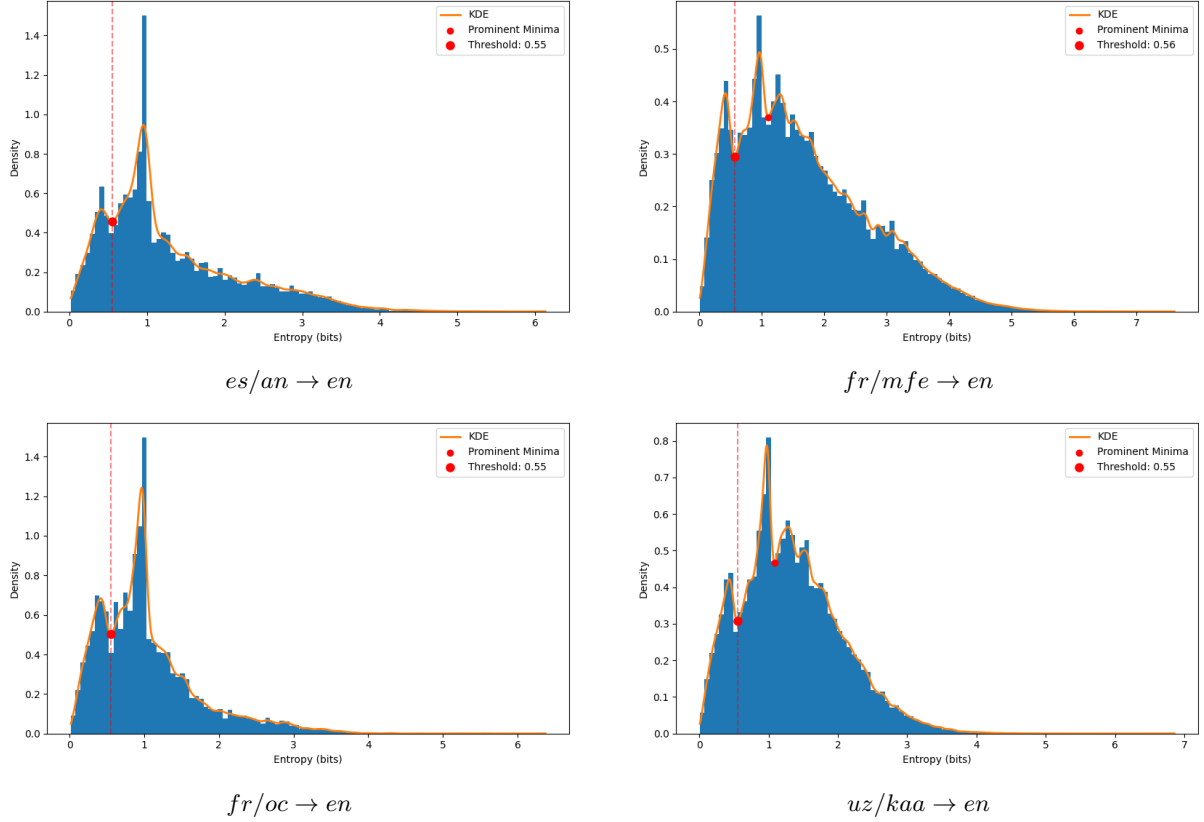


Figure 4: Distribution of n -gram correspondence entropies between PL and PL' after applying the respective PL/CL OC model to the $PL \rightarrow TL$ NMT training data with KDE-based entropy thresholds at the valley between the first two modes.

Model	n Samples	MaxSteps	BatchSize	LR	Wt.Decay	Val./epoch	Patience
$es/an \rightarrow en$							
Parent	10M	350K	512	5e-4	0.01	4	10
Child	20K	20K	32	5e-5	0.0	2	5
Simple	20K	20K	32	5e-5	0.0	2	5
$fr/mfe \rightarrow en$							
Parent	10M	350K	512	5e-4	0.01	4	10
Child	19K	20K	32	5e-5	0.0	2	5
Simple	19K	20K	32	5e-5	0.0	2	5
$fr/oc \rightarrow en$							
Parent	10M	350K	512	5e-4	0.01	4	10
Child	1.5M	150K	512	5e-5	0.01	4	10
Simple	1.5M	200K	512	3e-4	0.01	4	10
$uz/kaa \rightarrow en$							
Parent	10M	350K	512	5e-4	0.01	4	10
Child	51K	20K	32	5e-5	0.0	2	5
Simple	51K	20K	32	5e-5	0.0	2	5

Table 18: Training configurations for all NMT models. Parent refers to the pretrained models trained on either $PL \rightarrow CL$ (transfer-learning baseline) or $PL' \rightarrow CL$ (CharLOTTE). Child refers to the fine-tuned models trained on either $CL \rightarrow TL$ (transfer-learning baseline) or $CL' \rightarrow TL$ (CharLOTTE). Simple refers to the simple baselines trained on $CL \rightarrow TL$ without parent-child transfer learning. **n Samples** is the number of parallel sentences in the training set. **BatchSize** is the *effective* batch size.

Parameter	Value
Layers	6
Attention Heads	8
d_{ff}	2048
d_{model}	512
Max position embeddings	512

Table 19: Architectural configurations for all NMT models. Layers is the number of layers in both the encoder and decoder. Parameters not mentioned are kept at their default BART values.

C.2 Model Configuration and Training

All proposed and baseline NMT models are trained with the BartForConditionalGeneration (Lewis et al., 2020) transformer architecture. We modified the model configurations as reported in Table 19 based on settings of transformer-base (Vaswani et al., 2017). Architectural parameters not mentioned were kept at their default BART values. When training, we applied gradient clipping and linear warm-up for $\frac{1}{20}$ of the max training steps followed by linear decay. We trained to convergence, minimizing validation loss and applying early stopping. Selection criteria of the top 10 checkpoints was validation BLEU. All other training hyperparameters are reported in Table 18, which outlines those for parent and child models for the CharLOTTE and transfer-learning baseline, and those for the simple baseline. We trained using NVIDIA A100 GPUS, with parent models on 4 GPUs over 2-3 days, $oc \leftrightarrow en$ child and simple models on 4 GPUs up to 1 day, and $an \leftrightarrow en$ and $mfe \leftrightarrow en$ child and simple models on 1 GPU under an hour. At inference, sequences were generated with a max length of 512 tokens.

D Vocabulary Overlap Metrics

Jensen-Shannon divergence (JSD) between probability distributions P and Q is defined as:

$$JSD(P\|Q) = \frac{D(P\|M) + D(Q\|M)}{2} \quad (2)$$

where D is the KL divergence and M is the point-wise mean of P and Q . The Jaccard similarity (J) of unique token sets A and B is defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (3)$$

Model	Evaluator 1			Evaluator 2			Evaluator 3			Evaluator 4		
	Rk.	Win	Tie	Rk.	Win	Tie	Rk.	Win	Tie	Rk.	Win	Tie
<i>es/an</i> → <i>en</i>			47			78			97			87
CharLOTTE	1.42	112		1.41	88		1.53	75		1.18	78	
Transfer	1.78	41		1.68	34		1.77	28		1.39	35	
<i>uz/kaa</i> → <i>en</i>			46			32			18			44
CharLOTTE	1.34	111		1.34	115		1.39	122		1.39	118	
Transfer	1.68	43		1.65	53		1.70	60		1.79	38	

Table 20: Individual evaluator results for the human rank evaluation. **Rk.** ($\downarrow$) = rank metric, **Win** ($\uparrow$) = number of wins, **Tie** = number of ties.

Index	PL <i>n</i> -gram	Frequency Distribution of Mappings	Entropy (<i>H</i>)
1	gánd	gand:218	0.00
84	añam	anyam:24	0.00
196	\$580	\$5.80:15	0.00
287	rojo-	royo-:13	0.00
288	semá	sema:13	0.00
449	gír	chir:10	0.00
515	imét	imet:63, impat:1	0.12
532	jempl	xempl:95, jempl:2	0.14
579	üí	üi:67, üí:2	0.19
623	temá	tema:83, temá:3	0.22
735	jerar	herar:20, ferar:1	0.28
790	prej	prech:35, preix:2	0.30
831	dicán	dican:33, dicán:2	0.32
863	ducié	duci:46, ducie:3	0.33
978	treo\$	treu\$:14, treo\$:1	0.35
997	áli	ali:199, áli:15	0.37
1035	pnó	pno:13, pnó:1	0.37
1049	nados	natos:479, nados:34, ns:1	0.37
1092	\$485	\$4.85:12, \$4,5:1	0.39
1196	alíes	alís:11, alis:1	0.41
1268	poló	poloc:21, poló:2	0.43
1446	2/08/	2/08:9, 2/0/:1	0.47
1549	díge	diche:23, e:1, dige:1	0.48
1554	unk>.	.:39, 1:1, l:1, ulgar:1	0.48
1607	giz	chiz:129, giz:16	0.50
1639	nimá	nima:16, nimá:2	0.50
1661	dicié	dici:31, dicié:2, dicie:1	0.51
1701	cáse	case:15, cáse:2	0.52
1731	jert	chert:20, jert:1, hert:1	0.53
1809	autó	auto:47, autoc:3, autó:2	0.55

Table 21: Sample of meaningful *es*, *es'* *n*-gram mappings from scenario *es/an* → *en*. \$ = word boundary.

Index	PL n -gram	Frequency Distribution of Mappings	Entropy (H)
1	\$ent	\$ant:950	0.00
71	\$Cir	\$Sir:144	0.00
303	Sacr	Sakr:67	0.00
733	Tik	tik:41	0.00
835	abyl	abil:38	0.00
930	ferro	fero:36	0.00
1004	ttaw	tal:35	0.00
2069	dra-t	dret:22	0.00
3077	ascad	askad:17	0.00
3351	Necr	Nekr:16	0.00
3865	mochi	mosi:15	0.00
4404	si-o	zi-o:14	0.00
6881	dB1	del:10	0.00
7498	\$vé	\$ve:326, \$v:1	0.03
8360	attra	atra:135, atr:3	0.15
8729	iqua\$	ik\$:38, ike\$:1	0.17
10403	\$DG	\$De:90, \$Dek:1, \$Den:1, \$Dez:1	0.26
12441	thiop	tiop:40, Ø:1, tisp:1	0.32
12775	g1\$	g\$:15, ge\$:1	0.34
14182	éo-P	eo-P:13, eode:1	0.37
14596	Akhi	Aki:12, A:1	0.39
14730	eDr	er:12, err:1	0.39
15476	Della	Dela:11, Del:1	0.41
15988	oser\$	oze\$:116, oz\$:3, ozer\$:2, ose\$:1, use\$:1	0.42
16236	mapp	map:27, ap:1, Ø:1	0.43
17743	\$Frit	\$Frit:9, \$Fri:1	0.47
18653	Samb	Sanb:44, sanb:5	0.48
19372	tigu	tig:50, tigi:4, tigw:1	0.51
19906	Toto	toto:20, oto:1, to:1	0.53
20702	anni	ani:395, a:15, ni:5, anni:5, Ø:5, anii:1, ...	0.56

Table 22: Sample of meaningful fr , fr' n -gram mappings from scenario $fr/mfe \rightarrow en$. \$ = word boundary. Ø = empty string.

Index	PL n -gram	Frequency Distribution of Mappings	Entropy (H)
1	\$amé	\$ame:313	0.00
114	\$illé	\$ille:30	0.00
148	oDi	odi:27	0.00
149	xoné	xone:27	0.00
233	toté	tote:21	0.00
349	émemb	ememb:17	0.00
377	nk>q	Ø:16	0.00
650	J'ou	j'ou:11	0.00
694	ntisy	ntisi:11	0.00
871	éromè	eromè:10	0.00
1032	iSe	ise:32, Ise:1	0.20
1100	nTr	ntr:27, nTr:1	0.22
1264	flète	flete:20, flète:1	0.28
1480	sHe	she:16, sHe:1	0.32
1493	érea	erea:16, eraa:1	0.32
1525	erF	erf:60, erF:4	0.34
1613	\$assé	\$asse:14, \$assé:1	0.35
1766	tizat	tizac:13, tizat:1	0.37
1824	\$BitC	\$Bitc:12, \$BitC:1	0.39
1906	i-fé	i-fe:23, i-fé:2	0.40
2058	évide	evide:32, évide:3	0.42
2087	raël\$	raël\$:86, rael\$:6, \$:1	0.43
2230	éphar	ephar:10, éphar:1	0.44
2382	hWi	hwi:9, hWi:1	0.47
2483	gP	gp:46, gP:2, GP:2	0.48
2487	raré	rare:23, rare:1, rarat:1	0.48
2617	alD	ald:31, alD:4	0.51
2745	énot	enot:46, enota:4, not:1	0.53
2792	amC	amc:14, amC:2	0.54
2839	eV	ev:180, eV:14, EV:4, Ø:1	0.55

Table 23: Sample of meaningful fr , fr' n -gram mappings from scenario $fr/oc \rightarrow en$. \$ = word boundary. Ø = empty string.

Index	PL n -gram	Frequency Distribution of Mappings	Entropy (H)
1	ktsi	kci:1361	0.00
217	'qon	qan:79	0.00
1075	'mbo	mba:28	0.00
2138	'diyo	diya:17	0.00
2469	\$durl	\$dúrl:15	0.00
3143	ling'	limg':13	0.00
3841	dkag	dkağ:11	0.00
3900	jodul	jadıl:11	0.00
4251	Oysa	Aysa:10	0.00
6063	\$qadd	\$qádd:33, \$qadd:1	0.19
6508	tugil	túyil:67, tugil:1, tuwil:1	0.22
7689	Qayoq	Qayaq:19, Qaya:1	0.29
7706	boyin	bayın:19, boyın:1	0.29
7768	zonil	zanıl:19, zonıl:1	0.29
7806	irkat	irket:68, irkat:2, ırqat:1	0.29
8865	hbank	bank:15, hbank:1	0.34
8918	oz'l	azl:15, ozl:1	0.34
9007	Yara	Jara:397, Yara:17, YArA:1, ara:1, a:1, ılara:1	0.34
9026	bitir	pitir:330, bitir:20, pitke:1	0.34
9518	\$kurg	\$kúrg:65, \$kurg:3, \$kurg':1	0.37
9711	hnayo	naya:13, hnayo:1	0.37
9896	onats	onac:38, anac:3	0.38
9924	\$botl	\$batl:25, \$botl:2	0.38
10253	koysa	keyse:12, koysa:1	0.39
10472	zayti	zaytı:142, zayti:6, zeyti:2, zaytt:1	0.40
11751	iksga	ikske:10, iksqa:1	0.44
13706	\$Qilg	\$Qılğ:62, \$Qilg:8	0.51
13723	'hton	htan:21, hton:1, tan:1	0.51
14108	varig	varın:20, warın:1, irin:1	0.53
14530	Toz	Taz:69, Toz:5, coz:1, Tazı:1	0.55

Table 24: Sample of meaningful uz , uz' n -gram mappings from scenario $uz/kaa \rightarrow en$. \$ = word boundary.

Quality Assessment

The following sentence has been translated 2 times. Please read the sentence and its translations and then rank them by entering a number from 1 to 2, where 1 is the best. If the translations are equivalent, feel free to put the same number for them.

Content theories are centered on finding what makes people tick or appeals to them.

The theories of speech focus on seeking what attracts or has pleased people.

The theories of articles focus on seeking what attracts or delights the Devil.

Previous

Sentence 1 of 100

Next

Sentence	Complete
<u>1</u>	✗
<u>2</u>	✗
3	✓

Figure 5: The human rank evaluation user interface. For each sample, the order in which the machine translations are presented is shuffled without any indication of the system that generated them.

37086