Probing Multilingual Cognate Prediction Models

Clémentine Fourrier
Inria, France
clementine.fourrier@inria.fr

Benoît Sagot
Inria, France
benoit.sagot@inria.fr

Abstract

Character-based neural machine translation models have become the reference models for cognate prediction, a historical linguistics task. So far, all linguistic interpretations about latent information captured by such models have been based on external analysis (accuracy, raw results, errors). In this paper, we investigate what probing can tell us about both models and previous interpretations, and learn that though our models store linguistic and diachronic information, they do not achieve it in previously assumed ways.

1 Introduction

In historical linguistics, **cognates** are words that share a common etymological origin in a common parent language. Galician, Portuguese and Spanish *gato*, Catalan and Occitan *gat*, Italian *gatto*, French *chat* and Aromanian *cãtushi*, all meaning 'cat', as well as Romanian *cătușă* 'manacle',[1] are cognates, as they all descend from the same word *cattus* 'cat' in their mutual parent language, Latin. The parent word form *cattus* is called the **proto-form**. Comparing the phonetic form of sets of cognates allows to identify patterns: in our example, initial [g] in Galician to Italian corresponds to [ʃ] in French and [k] in Romanian and Aromanian. If said pattern is attested in more cognate sets, it is then considered to be a **sound correspondence pattern**, which emerge in related languages from the application of minimal, regular and exceptionless **sound changes rules** to the ancestral proto-forms.[2] Such sound correspondence patterns then help finding new cognates.

The **cognate prediction task** aims at predicting, from a phonetised word, the plausible phonetic form of its cognate in a related language, according to known sound correspondence patterns; this has many applications, from identifying new words with field linguists (Bodt et al., 2018; Bodt and List, 2019) to inducing translation lexicons for low-resourced languages (Mann and Yarowsky, 2001).[3]

This task has been modelled as a sequence to sequence character level machine translation task in the most recent papers studying it (see the survey on cognate prediction in Dekker and Zuidema (2021)), which drew linguistic conclusions on the latent information learnt by such models by studying their outputs in a 'black-box' fashion. However, no paper that we know of tried to confirm or inform these conclusions by using modern interpretability tools, such as probing tasks, hidden representation analysis, or inner components analysis.

In this paper, we therefore investigate whether the linguistic conclusions previously reached 1) can be reproduced, 2) hold under the scrutiny of modern interpretability techniques, and 3) can be extended. We first train several neural cognate prediction models,[4] and analyse their outputs as such. Then, we focus on applying modern interpretability techniques, and compare the insights they provide with prior hypotheses.

2 Related Works

2.1 Automatic Cognate Prediction

Automatic cognate prediction has been studied using character-level machine translation techniques (Beinborn et al., 2013; Wu and Yarowsky, 2018; Dekker, 2018; Hämäläinen and Rueter, 2019; Four-

[1] In Aromanian and Romanian, the words also underwent diminutive suffixes (*-ushi* and *-ușă*) additions to the now lost cognate root.

[2] These sound changes are assumed to be regular and without exception since (Osthoff and Brugmann, 1878), who stated that 'Every sound change [...] takes place according to laws that admit no exception'.

[3] Inferring the plausible shape of the related proto-form from its children (proto-form reconstruction) can be seen as a sub-task of cognate prediction.

[4] Training can be replicated using data provided with the paper, and code at github.com/clefourrier/CopperMT. We can provide all our trained models on request (>10GB).

Findings of the Association for Computational Linguistics: ACL 2022, pages 3786 - 3801
May 22-27, 2022 ©2022 Association for Computational Linguistics

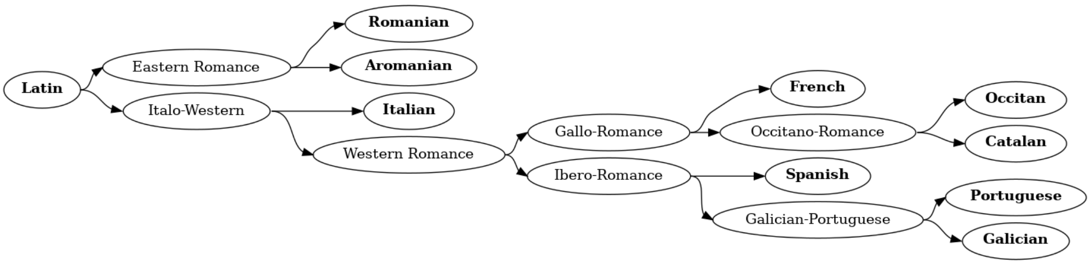

Figure 1: Relations between studied languages and their families.

rier and Sagot, 2020a). Dekker and Zuidema (2021) provide an overview of the different neural approaches used to solve this task (including their own), as well as its applications to other historical linguistic tasks (such as phylogeny reconstruction). However, the current paper follows specifically the tracks of two previous works studying encoder-decoder models for Romance cognate prediction.

Fourrier et al. (2021) study which NMT architecture fits the cognate prediction task best, comparing different methods and data augmentation techniques. They conclude that best results are obtained with multilingual RNN encoder-decoders with attention, a setup we shall follow. Meloni et al. (2021) train an encoder-decoder on the prediction of Latin proto-forms from modern Romance cognates sets. They then settle to explain the results linguistically in a 'black-box' fashion; we shall probe their conclusions.

2.2 Neural Models Interpretability

NLP interpretability is a recent field, with the first workshop dedicated to the topic occurring in 2018 (BlackBoxNLP, colocated with EMNLP 2018). Madsen et al. (2021) provide a review of post hoc interpretability techniques (focused on a posteriori model interpretation), which they divide along the level of abstraction (from local to global explanations). Among all the works they mention, we focus on two. Belinkov et al. (2020) develop toolkits for global interpretability in their tutorial: probing tasks and model components interaction and visualisation. Conneau et al. (2018) focus on probing tasks for sequence to sequence models, to investigate different aspects of language captured by the model. In this paper, we therefore focus on global post hoc interpretability techniques, such as visualisation and probing tasks, to linguistically interpret our models.

3 Paper Objective

3.1 Reference Task: Cognate Prediction

Training Objective The task we are optimising for is cognate prediction, i.e. generating, from a phonetised word, the plausible phonetic forms of its cognates in related languages. This is a sequence to sequence translation problem, going from a sequence of phones to a sequence of phones. To evaluate such 'translations,' we use Post (2018) implementation of BLEU (Papineni et al., 2002), which does not suffer for cognate prediction from the same drawbacks as for NMT (Fourrier et al., 2021).

Reference Architectures Best performing models for the task are NMT encoder-decoder models (Fourrier et al., 2021). They are composed of one or several encoder components, encoding the source word into a hidden representation, and of one or several decoder components, each playing the role of a 'conditional language model' (Conneau et al., 2018) that generates the output, in our case the target phonetic form of the word.

Languages Choice Sound correspondences and sound change rules are identified by looking at multilingual sets of cognates. If we want our neural models to latently capture such linguistic information, we need our data to be as multilingual as possible in a given language family.

We select 9 related Romance languages for which enough cognate data is available: Galician (GL), Portuguese (PT), Spanish (ES), Catalan (CA), Occitan (OC), Italian (IT), French (FR), Romanian (RO) and Aromanian (RUP).

The Romance family divided early in two branches (Fig. 1): the Eastern Romance branch (RO, RUP), and the Italo-Western branch (all others). They therefore constitute the two oldest language clusters in our data. However, through external influences on their phonology, French (Germanic influences) and the Eastern Romance branch (Slavic influences) tend to diverge from the other Romance languages studied. At the opposite end in the spectrum in terms of language closeness, Portuguese and Galician belong to their own language sub-branch, the Galician-Portuguese branch, as do Catalan and Occitan in the Occitano-Romance branch.

3.2 Steps of Analysis

We will first analyse our models and try to understand what they learned based only on their raw scores and prediction errors, as was done by Fourrier et al. (2021) and Meloni et al. (2021), to see the amount of linguistic information we can extract as such.

Then, we will probe the models, in order to compare the insights we got from a 'black box' analysis to insights obtained when probing specifically for linguistic or historical information. We therefore design the following probing tasks.

3.2.1 Synchronic Probes

Cognates are representative of their language phonetics, and we want to study whether the models learn deeper linguistic information while training on them.

Phonotactics To study whether our models learn phonotactics (the allowed arrangement of sounds and sound patterns in a language),[5,6] we adapt the *bigram shift* probing task (Conneau et al., 2018) to test whether encoders are sensitive to legal phone orders. A binary classifier is trained to distinguish between hidden representations of normal words and words whose phones have been inverted.

Phonology To study whether our models learn phonologically meaningful representations, we study our high-dimensionality hidden representation for each item of our vocabulary, as suggested in Madsen et al. (2021). We reduce the dimensionality of our encoded representations using PCA (Pearson, 1901) and t-SNE (der Maaten and Hinton, 2008) and look at the emerging underlying organisation of the phonetic space, as was done in Jacobs and Mailhot (2019) and Shibata et al. (2020) for, respectively, seq2seq phonetic and LSTM syntactic representations analysis.

3.2.2 Diachronic Probes

Cognates carry the historical information of the evolution of their respective languages. We want to see how much of this information was explicitly learned by the model.

Sound Correspondences and Contextualised Changes Cognates are usually identified by sound correspondence sets, which they also help define (see Sec. 1). Meloni et al. (2021) provide sample sets containing minimal examples of sound correspondences, as artificial subwords in some Romance languages and the associated Latin parent. To see if our models learn these sound correspondences, we study if they can reconstitute these sets.

Proto-form Reconstruction Cognates descend from a common ancestor word, their proto-form. When a multilingual neural model learns mappings between cognates in related languages, the shared joint intermediate representation tends towards their common denominator.[7] A plausible candidate would be a mapping of a common ancestor space, as proto-form have the overall smallest distance to all their children. To study whether the model contains historical information about the proto-forms, we design a probing task where we train a decoder to predict a Latin word from the fixed encoded representation of its children Romance cognates.

4 Detailed Experimental Setup

4.1 Data

Extraction and Pre-processing Monolingual[8] and bilingual[9] cognate lexicons are extracted from `EtymDB2` (Fourrier and Sagot, 2020b), an etymological database, using the scripts provided. All data is then phonetised using `espeak` (Duddington, 2007-2015), with relevant phonetizers for CA, ES, IT, FR, PT, RO, and approximating the phonetization of OC as CA, RUP as RO, and GL as PT.[10] We segment the data at the character level then split it 85/7.5/7.5% for the train/dev/test sets (see

[5] e.g. In Spanish, a word can start with [est] but not [st].

[6] Phonotactics, in a sense, is the 'syntax' of phonology.

[7] As each encoder must store information for all decoders, and each decoder read information from any encoder, the multilingual intermediate representation converges.

[8] Our monolingual cognate lexicons contain words that descend directly from our languages' common ancestors and are likely to belong to cognacy relations.

[9] Bilingual cognate lexicons contain attested cognate pairs.

[10] These approximations should hold for our study, as these languages have the most linguistic features in common.

App. A.1.2). The split is repeated 3 times with different shufflings for statistical significance.

Description There is considerable variability in the number of word pairs between our bilingual datasets (see Appendix, Table 5): OC→RUP (two of our least resourced languages) contains 81 pairs, whereas PO→ES contains 1930 pairs. Monolingual datasets vary from 553 words for OC to 6005 words for IT, with CA, ES, FR, IT, and PT sets containing more than 2000 words, and GL, OC, RO and RUP less than 1500.[11] The total number of phones per pair varies accordingly; the number of unique phones per language pair stands between 32 and 56, depending on the number of shared phones between languages. Average word length varies between 5.3 and 8.3 phones.

4.2 Models

Name	#source	#target	With mono	Sharing
SMT	1	1	No	-
Bi-NMT	1	1	No	None
Bi-NMT+m	1	1	Yes	None
M-NMT	9 (all)	9 (all)	No	None
M-NMT+m	9 (all)	9 (all)	Yes	None
+shared_emb	9 (all)	9 (all)	Yes	Embeddings
+shared_all	9 (all)	9 (all)	Yes	All

Table 1: Model type setups

The summary of all our encoder-decoder models is developed in Table 1. Our baselines are SMT models trained for each language direction (**SMT**), more adapted to very low-resource setups. We train bilingual NMT models, without (**Bi-NMT**) or with (**Bi-NMT+m**) added monolingual data,[12] and multilingual models without (**M-NMT**) or with (**M-NMT+m**) monolingual data, using one encoder and one decoder per language. We also study the impact of sharing components in our likely best setup (in terms of data size seen by the model: M-NMT+m), and either share embeddings layers (**M-NMT+m+shared_emb**) or share full encoders and decoders across all languages (**M-NMT+m+shared_all**). Training details can be found in Appendix A.2.

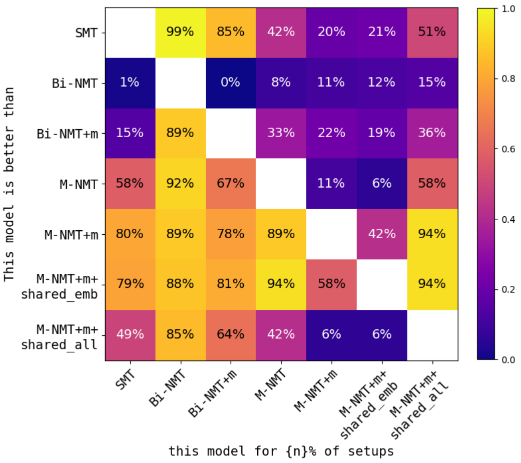

Figure 2: Percentage of language pairs for which a given model (left) outperforms an other (bottom).[13]

5 Blackbox Analysis

5.1 Raw BLEU Results

The full BLEU score tables of all our models on all our language pairs are in Appendix A.5.

5.1.1 Best Setup Choice

We synthesise the respective performance of our models in Fig. 2, comparing their BLEU scores. This heatmap indicates the percentage of language pairs for which a model (left) is better than another model (bottom). Both Bi-NMT models perform worse than the SMT baseline (with and without monolingual data). Multilinguality improves the performance, as the M-NMT model outperforms the baseline in 58% of cases. However, the best results are obtained when the models see the most data; the different M-NMT+m models outperform all other models for 80% of language pairs minimum. Another slight increase is obtained by sharing embeddings, as the M-NMT+m+shared_emb outperforms the M-NMT+m model in 58% of cases. We will therefore focus on the M-NMT+m and M-NMT+m+shared_emb models, our two best setups.

5.1.2 Impact of Parameters

To study performance on all language pairs separately, we generate the heatmap of average BLEU scores (Fig. 3) from all sources (y-axis) to all targets (x-axis) for our two best architectures and the baseline, with high/low scores in red/blue, and big/small datasets indicated by +/− respectively. Our models and baseline behave similarly, with

[11] We use monolingual data to reinforce the decoders language modelling capabilities, see next section. We expect that such a variation in size will impact learning.

[12] Bi-NMT+m models train on a single language pair, augmented with the monolingual target data, provided to the decoder through its own encoder; they allows the target decoder to see as much target data as possible, to reinforce it language modelling capacities.

[13] Sums not equal to 100% indicate that the models have the same performance on some language pairs (ex: Bi-NMT and Bi-NMT+m).

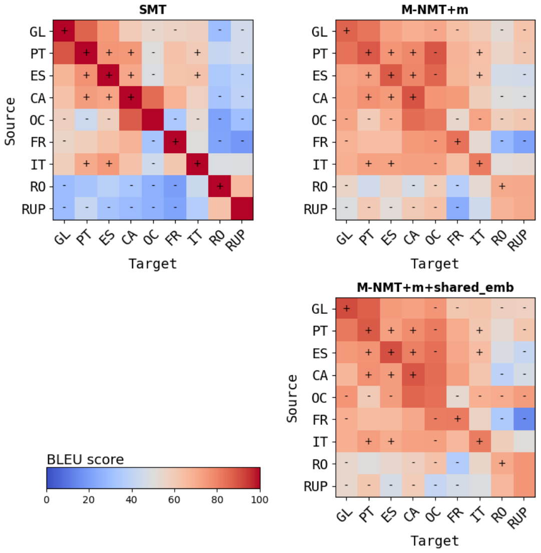

Figure 3: Heatmap of the BLEU scores for our models of interest.
Languages: source in y, target in x. Data size: + indicates more than 1000 word pairs, − less than 300.

overall good BLEU scores, which seem to be slightly correlated with data size, except for some outliers. Firstly, predicting RO and RUP from/to all other languages has a considerably lower BLEU than all other pairs, except for RO–RUP itself: predicting between languages from too dissimilar language branches (Eastern-Romance and Italo-Western Romance), unsurprisingly, seems harder than translating within either of those branches. Secondly, GL↔PT and OC↔CA have higher BLEU than we could expect based on data size only.[14] In all setups, it therefore appears to be easier to predict cognates for closely related languages.[15]

5.2 Predictions Analysis

We compare the predictions and errors made by the models in three cases: the language pair is highly resourced and gets a good BLEU score (ES-PT), the language pair has average resources but contains close languages and gets a good BLEU score (PT-GL), the language pair has almost no resource and gets a bad BLEU (RO-FR).

We use the Needleman and Wunsch (1970) dynamic programming algorithm, modified by Gotoh (1982)[16] to compute the pairwise alignment between predictions and gold targets in 1 or 2-grams.[17] We can then better see which predicted phones match the gold or not, and why.[18]

5.2.1 General Observations

When looking at the phone level model predictions, we observe that they can be: (1) correct (equal to gold); (2) phonetically close to the gold (ex: [β], a voiced bilabial fricative, instead of [b], a voiced bilabial plosive); (3) either a known sound correspondence, incorrect in the current example but attested in others (ex: [v], a voiced labiodental fricative, instead of [b], a voiced bilabial plosive) or a wrong prediction (ex: [a], a vowel, instead of [b], a consonant) (Table 2). In 2-gram, this classification becomes (1) correct (identical 2-grams); (2) close (identical/close phone and close phone); (3) the rest, which can then be divided in (a) 'one correct/close and one wrong', or (b) 'two wrong' phones, other patterns almost not occurring.

For our high-resource pair (ES→PT), our models perform similarly to the baseline: they are correct in 90% of cases, and more often close than wrong the rest of the time.

We observe two different behaviours for our comparatively less-resourced pairs. For the pair with close languages (PT→GL), multilinguality decreases performance (by 2 to 5 points) with respect to the baseline. For our extremely low-resourced and sparsely related pair (RO→FR), however, the multilingual models outperform the SMT baseline for the first time (by 9 to 15 points), likely thanks to data augmentation provided by multilinguality. Sharing embeddings seems to have a significant impact only when the languages are far away and the data quantity low, as it inverts the ratio of close to wrong results from 1:3 to 3:2, seemingly increasing the model language modelling capability.

5.2.2 Error Patterns

Errors can be separated between those which occur only once, and tend to be nonsensical, and those

[14] It is important to note that this could also be linked to similarities introduced by our phonetisation method, as we used the Catalan phonetizer for Occitan and the Portuguese phonetizer for Galician.

[15] We can also observe than the diagonal - predicting from a language to itself - has lower score for M-NMT: using multilingual models tends to lower the accuracy when going from one language to itself, most likely because the "conditional language modeling" decoder for a given language is perturbed by noise introduced in the intermediate representation space when learning on other input languages.

[16] We use the BioPython (Cock et al., 2009) implementation.

[17] Using 3-grams alignments provided no further insights.

[18] To remove noise which might be caused by incorrect alignments, we only keep correspondences occurring more than once, and in 2-grams, we discard the pairs which contained a blank inserted during the alignment process.

Pair		ES→PT			PT→GL			RO→FR		
Prediction		Correct	Close	Wrong	Correct	Close	Wrong	Correct	Close	Wrong
1-gram:	SMT	90.9%	5.3%	3.8%	95.5%	2.4%	2.1%	62.7%	12.7%	24.5%
	M-NMT+m	89.1%	5.5%	5.4%	93.6%	3.4%	3.1%	71.6%	8.8%	19.6%
	+shared_emb	90.7%	5.1%	4.3%	92.4%	3.9%	3.7%	73.8%	14.6%	11.7%
2-gram:	SMT	83.4%	9.7%	6.9%	93.2%	4.1%	2.6%	49.1%	14.0%	36.8%
	M-NMT+m	81.5%	9.8%	8.6%	89.3%	6.2%	4.5%	64.4%	8.5%	27.1%
	+shared_emb	83.3%	9.6%	7.1%	88.0%	6.8%	5.2%	58.6%	24.1%	17.2%

Table 2: Prediction types frequency for 1 and 2 grams, for three language pairs: ES→PT (good BLEU, big data size), PT→GL (good BLEU, average data size, close languages), RO→FR (bad BLEU, small data size).

with a higher apparition frequency, which tend to be plausible and similar between neural models and baseline. We only analyse frequent errors in the following section, therefore not studying RO→FR, whose errors tend to occur only once and be nonsensical (likely the result of the difficulty of learning on so little data).

Wrong phones in 1-gram or 2-gram case (a) correspond to high-mid vocalic alternations patterns, ([ɔ]/[u], [ɛ]/[ɨ]-[i]), exchange of consonants linked by a sound correspondence ([v]/[b]), or less frequently, in 2-gram only, to a [k]/[ʒ] or [w]/[l] confusion.[19] 2-gram case (b) correspond to metathesis (phone inversions, ex: [ɪŋ]/[nɨ] or [eɾ]/[ɹɨ]) 30% of the time, the rest being nonsensical errors.

These results seem to confirm the observations made by Meloni et al. (2021) that most errors made by the models are not arbitrary but tend to correlate with historical linguistic phenomenon.

5.3 Conclusion

Analysing our models using standard error analysis methods allow us to conclude that (1) multilinguality helps considerably to predict cognates, which might reflect information transfer or sharing in the models, and (2) errors made by the models suggest that they learn (a) phonetic similarity and (b) linguistic phenomena.

6 Synchronic Probing

Using previously defined probes, we study whether our models learn synchronic linguistic information.

6.1 Phonotactics

Probe Training We trained MLP classifiers to detect whether encoded words contain a switched bigram of phones or not. For a given language, the encoder used is either randomly initialised or coming from our multilingual models. This experiment is reproduced for all data shuffles and all languages. No matter the setup, the classifier performance is systematically around 50%, no better than random.

Fine-tuning We decide to try fine-tuning our multilingual models on the classification of bigram switches, to see if this is information our models can learn to distinguish. We use the same setup as for the probing tasks, except that the encoders are now fine-tuned along the classifier training. The results are again no better than random.

Conclusion When learning to predict cognates, the encoder does not spontaneously encode phonotactics information, nor does it learn to encode it when fine-tuned specifically on that. This is interesting, because sound correspondences relations between cognates are partly linked to phonotactics. If the model does not learn this information explicitly, it has to learn something else instead.

6.2 Vocabulary Information

We study learned phone proximity by using dimension reductions techniques (PCA, t-SNE) on the encoders' hidden representation. We present here 3-dimensional PCA for the vowels' representations (Fig. 4), but observations we make also hold true for consonants (see Appendix A.4).

Language Relatedness Along one dimension, the space seems to be organised through a linguistic continuum (with vowels in French together, then the rest of the Gallo-Romance branch, then the Eastern-Romance branch, then the Ibero-Romance branch).[20] However, this continuum is not constant across data shufflings; depending on the data seed, the model places different languages close to one

[19] SMT also produce a segment voicing change between [ŋv] and [mb].

[20] Clustering phones on their respective languages is the main feature we observe when using t-SNE.

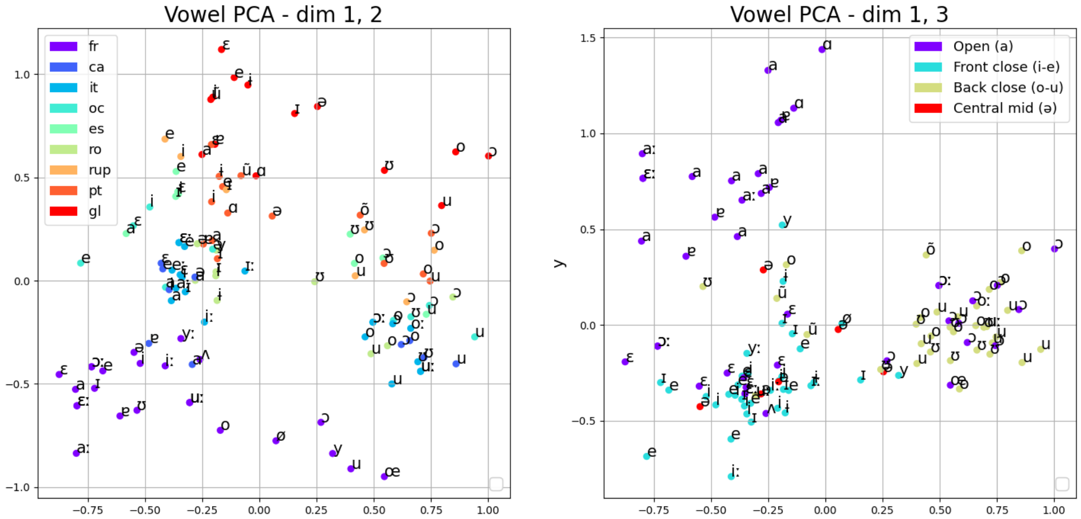

Figure 4: Vowels PCA, seed 0. Left: coloured on language. Right: coloured on pole of the vocalic triangle

another in the intermediate representation—models learn a language separation of the space, but not constant language relationships.

Phonetic Organisation Along the other two dimensions appears a pattern of phonetic organisation seemingly similar to the vocalic diagram,[21] which proves stable across all our runs. All our NMT models, no matter the data shuffling trained on, seem to have the three phonetic vocalic poles in their PCA ('u/o', 'i/e', and 'a'), more or less some outliers. These outliers fall in two categories: rare French phones (e.g. nasal vowels, which do not exist in the other Romance languages, and therefore are harder to place), or, interestingly, phones actually clustered with the most similar pole orthographically and not phonetically. For example, ɔ is linked to 'u/o' instead of 'a' (and both [ɔ] and [o] sounds usually come from the letter o), ɛ to 'i/e' instead of 'a' ([ɛ] and [e] from e). The models appear to have learned to encode similarly phones occurring in similar contexts, and not phones that are actually phonetically similar.[22] We can therefore say that, though the models *seem* to have learned a 'phonologically meaningful taxonomy of phonemes without explicit supervision' (Meloni et al., 2021), a faithful and not just plausible interpretation (Jacovi and Goldberg, 2020) is that they have actually learned something akin to a 'phonetic language model'. However, since sound changes occur regularly, phones in similar contexts in related languages will tend to have evolved from a common ancestor phone: closer intermediate representations belonging to contextually similar phones tends to confirm a form of historical mapping.

7 Diachronic Probing

7.1 Do the Models Learn Phone Correspondences?

Spanish to	IT	PT	FR	RO	Avg.
SMT	**76**	**73**	**64**	**73**	**71**
M-NMT+m	67	61	52	61	60
+shared_emb	61	61	58	64	61
Italian to	ES	PT	FR	RO	Avg.
SMT	**88**	64	**73**	**76**	**75**
M-NMT+m	61	**70**	27	58	54
+shared_emb	70	61	52	55	59
Portuguese to	ES	IT	FR	RO	Avg.
SMT	**88**	**82**	67	**76**	**78**
M-NMT+m	76	76	**76**	70	74
+shared_emb	73	67	55	67	65
French to	ES	IT	PT	RO	Avg.
SMT	61	67	36	**64**	57
M-NMT+m	70	**70**	**76**	61	**69**
+shared_emb	**73**	64	**76**	48	65
Romanian to	ES	IT	PT	FR	Avg.
SMT	**72**	62	59	**62**	**64**
M-NMT+m	56	**69**	**66**	34	56
+shared_emb	53	**69**	62	41	56

Table 3: % of cases where our models predicted the good artificial correspondence among the 5-best predictions (for the Meloni et al. (2021) sets). Best in bold.

[21] The vocalic diagram is obtained when organising vowels along their production height and tongue advancement.

[22] However, phones occurring in similar contexts in our cognates usually come from the same original sounds, and therefore tend to be phonetically similar.

Meloni et al. (2021) provide sets of minimal phonemes test sequences representing known sound correspondences in RO, FR, IT, ES and PT, to evaluate their models' generalisation. For example, the minimal set for sound changes linked to word initial Latin /pl/ is, for an artificial Latin origin [pla]: RO [pla], FR [pla], IT [pja], ES [ʎa] and PT [ʃa]. We predict 5-best 'cognates' for the provided artificial segments, to see if our models can generalise sound correspondences too. The correct results appear in 1st or 2nd position most of the time (Table 6 in Appendix). Our neural models reach between 54% and 74% average accuracy from a given language (Table 3),[23] and the statistical baseline tends to perform better overall. However, sound correspondences where the source languages are the most divergent in our Romance family (French and Romanian, see Sec. 3.1) are better captured with the neural models by 3 to 40 points (for language pairs with enough data, such as FR→ES, IT, PT, or RO→IT, PT). Adding shared embeddings increases performance with our more typical Romance languages as source and decreases performance for the previous languages, while still performing better than the baseline. We can therefore say that sound correspondences information is captured by our models.

Model	CA	ES	FR
Top baseline	32.3 ± 4.7	**46.7** ± 0.6	**31.7** ± 3.6
M-NMT+m	**36.8** ± 1.3	38.8 ± 2.4	**31.7** ± 0.9
Bi-NMT+m	28.5 ± 3.7	38.0 ± 1.9	29.9 ± 0.8
Untrained baseline	5.2 ± 0.9	3.1 ± 0.5	3.1 ± 1.0
Model	GL	IT	OC
Top baseline	23.8 ± 4.3	**50.5** ± 3.0	6.5 ± 1.0
M-NMT+m	**26.8** ± 1.9	45.1 ± 0.6	**9.6** ± 1.4
Bi-NMT+m	20.7 ± 2.1	44.0 ± 0.6	9.0 ± 3.1
Untrained baseline	2.8 ± 0.5	5.5 ± 1.8	1.8 ± 0.1
Model	PT	RO	RUP
Top baseline	**36.4** ± 2.9	18.2 ± 6.2	9.9 ± 1.9
M-NMT+m	35.1 ± 0.6	21.1 ± 2.5	**18.1** ± 4.5
Bi-NMT+m	31.1 ± 0.9	**26.2** ± 0.8	16.8 ± 0.4
Untrained baseline	4.8 ± 0.7	2.6 ± 0.9	2.5 ± 0.3

Table 4: Probe BLEU test scores for 3 seeds (20 epochs)

7.2 Do the Models Capture Diachronic Information?

We used very small RNN decoders with attention[24] as probes, and trained them to predict Latin proto-forms from the NMT encoded hidden representations of several models. We trained our probes to predict from **M-NMT+m** frozen encoders. Then, to assess if multilinguality is helpful in capturing latent historical information, we trained probes on the source-to-source **Bi-NMT+m** frozen encoders, which have learnt a coherent hidden representation of the source language, but possess no extra linguistic information. To make sure that our probes are not too expressive, we trained some on an untrained encoder frozen after random initialisation, as an **untrained baseline** (Conneau et al., 2018; Zhang and Bowman, 2018). Too expressive networks can learn to fit any random noise, and have therefore no value as probes.[25] Lastly, we compare everything to the best possible setup, our **top baseline**: a Bi-NMT model trained specifically on the task of learning Latin from the current source. On Table 4, we plotted the BLEU test scores obtained at each epoch by the different setups for the different languages. Our bottom baselines' low performance confirms that our probes are selective enough to prevent rote memorisation of anything. M-NMT+m encoders, without any fine-tuning on the prediction of Latin, reach or surpass the performance of models specifically trained on this task, and are outperformed by our Bi-NMT+m encoders only once.[26] Multilinguality therefore introduces latent linguistic information, which helps reconstruct the proto-form better than when using bilingual models only.

8 Conclusion

After training and selecting the best multilingual machine translation models for the task of cognate prediction, we confirmed the black-box analysis previously made of similar models (they capture language relatedness information and phonetic similarity). We then probed our models and discovered that latent linguistic information learned by the model seemed to encode a phonetic 'contextual language model' rather than explicit phonology or phonotactics. We also discovered that our mod-

[23] We did not expect our models to reach a 100% accuracy, as the provided examples are minimal for a set, and not necessarily a sound pair between languages (some sounds could also appear in other sound correspondences), but reach nonetheless a comparable accuracy to (Meloni et al., 2021) on their similar proto-form prediction task.

[24] Embed./Hidden sizes: 10/20, Luong dot attention.

[25] 'As long as a representation is a lossless encoding, a sufficiently expressive probe with enough training data can learn any task on top of it' (Hewitt and Liang, 2019)

[26] The M-NMT+m+shared_emb encoders reach half the performance of the M-NMT+m model: sharing embeddings seems to capture considerably less diachronic information, possibly because the phonetic information of all languages are mashed together.

els learn diachronic information: they are able to produce sound correspondences, and, even more interestingly, they contain enough historical linguistic information to allow the reconstruction of the proto-form with no fine-tuning, performing at least as well as models trained specifically for this task. We can therefore conclude that synchronic multilingual cognate prediction models learn latent diachronic information, though further work is needed to understand more precisely under which form this information is stored.

Acknowledgements

This work was performed using HPC resources from GENCI-IDRIS (Grant 2021-AD011011459R1).

References

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.

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5, Online. Association for Computational Linguistics.

Timotheus A. Bodt, Nathan W Hill, and Johann-Mattis List. 2018. Prediction experiment for missing words in Kho-Bwa language data. *Open Science Framework Preregistration*.

Timotheus A. Bodt and Johann-Mattis List. 2019. Testing the predictive strength of the comparative method: an ongoing experiment on unattested words in Western Kho-Bwa languages. *Papers in Historical Phonology*, 4:22–44.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Peter Dekker. 2018. Reconstructing language ancestry by performing word prediction with neural networks. Master's thesis, University of Amsterdam.

Peter Dekker and Willem Zuidema. 2021. Word prediction in computational historical linguistics. *Journal of Language Modelling*, 8(2):295–336.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Jonathan Duddington. 2007-2015. espeak text to speech.

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 12th Language Resources and Evaluation Conference*, pages 3207–3216, Marseille, France. European Language Resources Association.

Osamu Gotoh. 1982. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162(3):705–708.

Mika Hämäläinen and Jack Rueter. 2019. Finding Sami cognates with a character-based NMT approach. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 39–45, Honolulu. Association for Computational Linguistics.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, UK. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Cassandra L. Jacobs and Fred Mailhot. 2019. Encoder-decoder models for latent phonological representations of words. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 206–217, Florence, Italy. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

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.

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.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2021. Post-hoc interpretability for neural nlp: A survey. *arxiv*.

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*.

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2021. Ab antiquo: Neural proto-language reconstruction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4460–4473, Online. Association for Computational Linguistics.

Saul B. Needleman and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Hermann Osthoff and Karl Brugmann. 1878. *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen*. Hirzel, Leipzig, Germany.

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.

Karl Pearson. 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

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.

Chihiro Shibata, Kei Uchiumi, and Daichi Mochihashi. 2020. How LSTM encodes syntax: Exploring context vectors and semi-quantization on natural text. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4033–4043, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Winston Wu and David Yarowsky. 2018. Creating large-scale multilingual cognate tables. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361, Brussels, Belgium. Association for Computational Linguistics.

A Appendix

A.1 Data Presentation

A.1.1 Data size

From Catalan (CA) to	CA	ES	FR	GL	IT	OC	PT	RO	RUP
#words	2612	1233	466	449	970	324	1031	235	144
#phones	16472	16171	5706	5724	12511	3486	13601	2162	1307
#unique phones	36	41	47	44	56	35	44	42	40
Avg word length	7.31	7.56	7.12	7.37	7.45	6.38	7.60	5.60	5.54
From Spanish (ES) to	CA	ES	FR	GL	IT	OC	PT	RO	RUP
#words	1236	4967	693	732	1880	230	1930	463	291
#phones	16198	34176	8931	9760	25686	2534	26156	4700	2898
#unique phones	41	35	46	44	54	38	44	42	39
Avg word length	7.55	7.88	7.45	7.67	7.83	6.51	7.78	6.08	5.98
From French (FR) to	CA	ES	FR	GL	IT	OC	PT	RO	RUP
#words	466	694	3772	215	715	110	600	135	86
#phones	5707	8941	21225	2641	9332	1126	7665	1183	737
#unique phones	47	46	46	42	54	41	43	37	36
Avg word length	7.13	7.44	6.63	7.15	7.53	6.12	7.39	5.39	5.30
From Galician (GL) to	CA	ES	FR	GL	IT	OC	PT	RO	RUP
#words	449	732	215	1464	558	138	882	176	106
#phones	5724	9759	2641	9509	7196	1455	11117	1703	1005
#unique phones	44	44	42	35	51	41	37	38	37
Avg word length	7.37	7.67	7.15	7.50	7.45	6.27	7.30	5.84	5.74
From Italian (IT) to	CA	ES	FR	GL	IT	OC	PT	RO	RUP
#words	973	1885	717	558	6005	234	1557	618	378
#phones	12534	25742	9346	7190	44073	2660	21199	6834	4046
#unique phones	56	54	54	51	49	50	55	50	47
Avg word length	7.44	7.83	7.52	7.44	8.34	6.68	7.81	6.53	6.35
From Occitan (OC) to	CA	ES	FR	GL	IT	OC	PT	RO	RUP
#words	324	230	109	138	234	553	222	117	81
#phones	3486	2534	1120	1455	2659	3026	2391	1044	724
#unique phones	35	38	41	41	50	33	42	38	36
Avg word length	6.38	6.51	6.14	6.27	6.68	6.47	6.39	5.46	5.47
From Portuguese (PT) to	CA	ES	FR	GL	IT	OC	PT	RO	RUP
#words	1031	1930	596	883	1556	223	4891	399	261
#phones	13606	26158	7624	11125	21188	2399	33046	3991	2569
#unique phones	44	44	43	37	55	42	37	39	38
Avg word length	7.60	7.78	7.40	7.30	7.81	6.38	7.76	6.00	5.92
From Romanian (RO) to	CA	ES	FR	GL	IT	OC	PT	RO	RUP
#words	236	465	136	175	621	117	398	1088	412
#phones	2173	4715	1193	1696	6859	1044	3984	5833	4251
#unique phones	42	42	37	38	50	38	39	32	32
Avg word length	5.60	6.07	5.39	5.85	6.52	5.46	6.01	6.36	6.16
From Aromanian (RUP) to	CA	ES	FR	GL	IT	OC	PT	RO	RUP
#words	146	292	87	107	378	81	259	412	817
#phones	1327	2907	745	1015	4038	724	2551	4251	4531
#unique phones	40	39	37	37	47	36	38	32	29
Avg word length	5.54	5.98	5.29	5.74	6.34	5.47	5.92	6.16	6.55

Table 5: Detailed dataset statistics for our lexicons.

A.1.2 Data segmentation and splitting

We segmented the data at the character (not subword) level using the `SentencePiece` (Kudo and Richardson, 2018) library; more precisely, we trained a character-level model per language for all models, except M-NMT+m+shared_emb and M-NMT+m+shared_all, where sharing embeddings or encoders meant sharing the vocabulary across all languages: in this last case, we used a single segmentation model for all languages (which tend to have similar phone distributions, apart from the rarest phones, such as nasal vowels in French). The vocab size parameter was 100, superior to the total number of unique phones.

As this is not a common task, there is no "standard" for splitting this kind of data set. We tried to balance training on the maximum amount of data possible (85%) without loosing accuracy (by asserting that our runs are statistically significant, launching all experiments with 3 different data splits).

A.2 Training Details

For our SMT baseline, we use the `Moses` toolkit to train an SMT model for each language direction. The data is aligned with `GIZA++` (Och and Ney, 2003), while a 3-gram language model is trained with `KenLM` (Heafield, 2011) on the pair of interest target data, then models are tuned using `MERT`.

For our NMT models, we use RNN encoder-decoder models with attention (Cho et al., 2014; Luong et al., 2015), since Transformers (Vaswani et al., 2017) have been shown to under-perform for this task because of data scarcity (Fourrier et al., 2021). We use the `fairseq` toolkit (Ott et al., 2019); the encoders are composed of one embedding layer followed by a bidirectional GRU (embedding dimension: 20, hidden dimension: 50, 1 layer), and the decoders are composed of one embedding layer and one unidirectional GRU with its own attention (same parameters). Each model can share encoders/decoders/embedding layers or not across languages. Each model is trained using the Adam optimizer (learning rate: 0.005) and the cross entropy loss, stopping on the first of either 15 epochs or convergence of the BLEU score on the development set used during training.

A.3 Sound Correspondence Prediction

We also compute the average position for the correct result among the 5-best predictions , and observe that all models have similar behaviours: when answers are correctly predicted, they usually are predicted in first or second position on average (the neural models being better than the baseline for our linguistically more original languages, Romanian and French).

Spanish to	Italian	Portuguese	French	Romanian
SMT	1.5	2.4	1.4	1.8
M-NMT+m	1.5	2.0	2.0	1.7
+shared_emb	1.2	1.4	2.0	1.8
Italian to	Spanish	Portuguese	French	Romanian
SMT	1.4	2.7	1.4	2.0
M-NMT+m	1.4	2.1	2.6	1.7
+shared_emb	1.7	2.0	1.9	1.9
Portuguese to	Spanish	Italian	French	Romanian
SMT	1.4	1.6	1.1	2.5
M-NMT+m	1.7	1.9	1.7	1.4
+shared_emb	1.5	1.9	2.3	1.7
French to	Spanish	Italian	Portuguese	Romanian
SMT	1.4	2.8	3.2	1.6
M-NMT+m	1.5	1.9	1.2	1.9
+shared_emb	1.6	1.9	2.0	2.2
Romanian to	Spanish	Italian	Portuguese	French
SMT	2.7	2.6	3.5	2.3
M-NMT+m	1.3	2.1	1.4	2.0
+shared_emb	1.2	1.9	1.7	1.8

Table 6: Average position of the correct result in 5-best

A.4 Consonants PCA and t-SNE

We plot the PCA (Figure 5) and t-SNE (Figure 6) for consonants, coloured on either manner or place, and observe the same patterns are mentioned in the paper. Letters seem to be grouped phonetically at a first glance, but are actually grouped by orthographic context more than phonetic similarity: ([b], [β], [v] together, or [g], [ɣ], [k] together, and so forth).

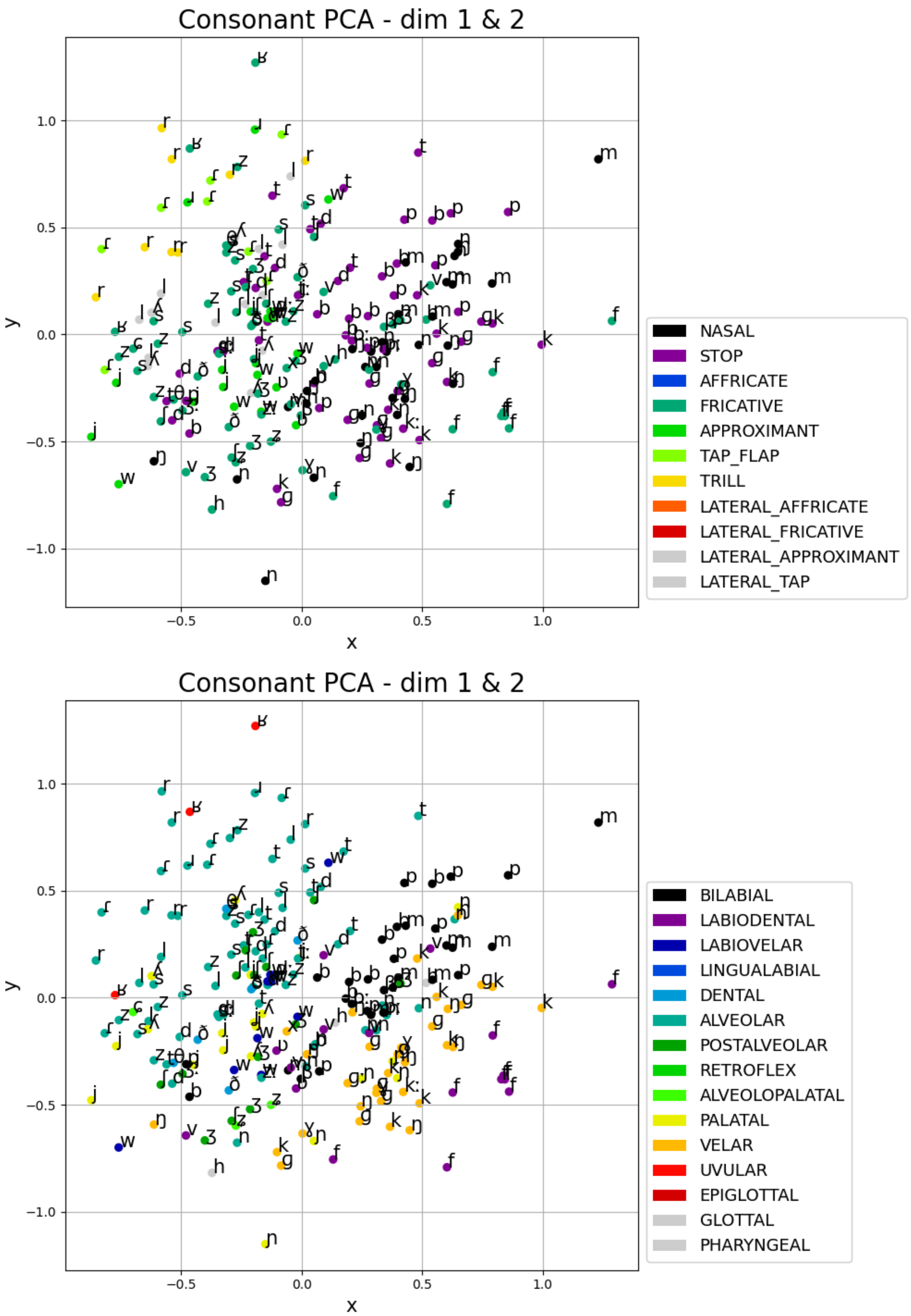

Figure 5: Consonant PCA, seed 0, coloured on manner above and on place below

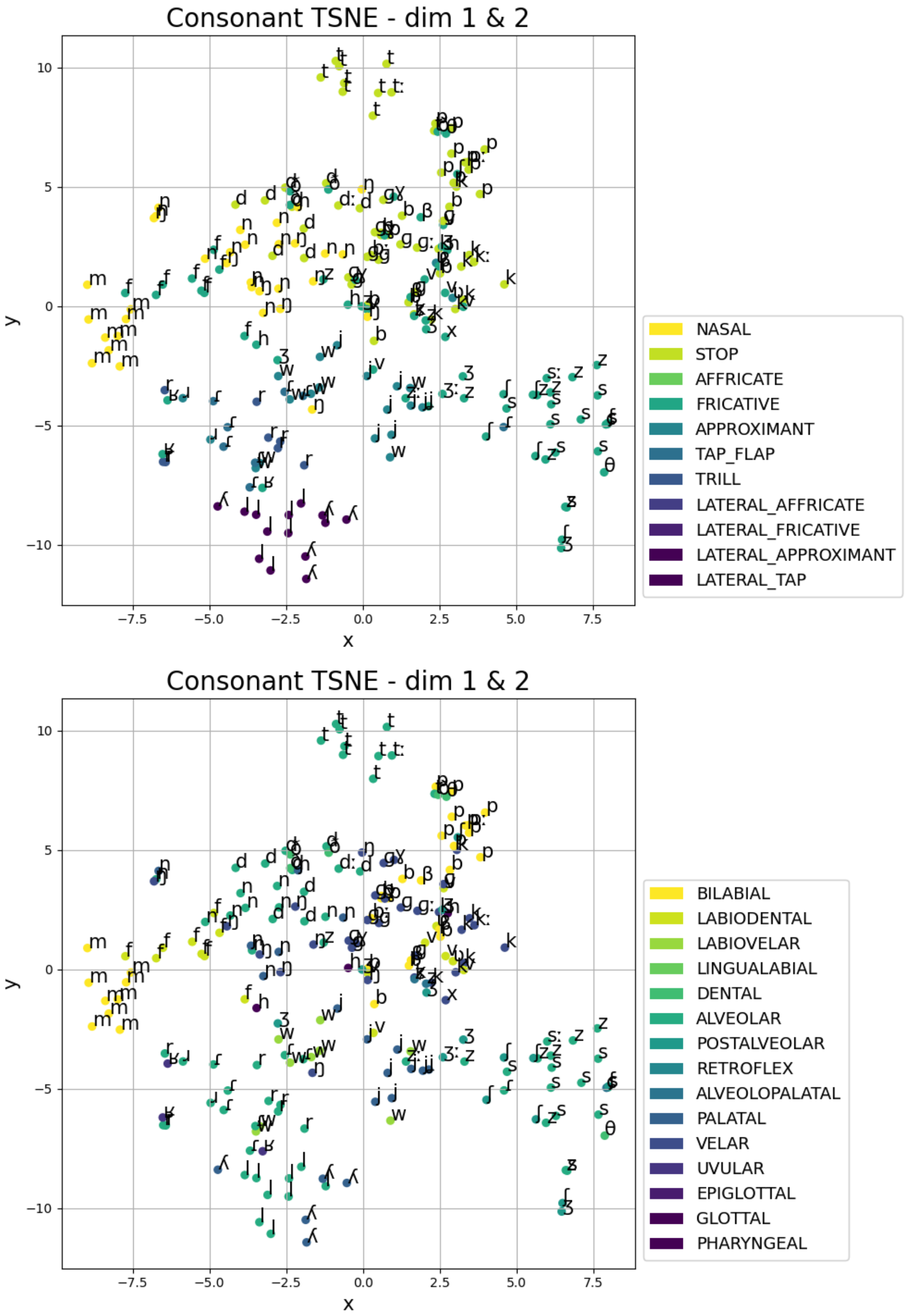

Figure 6: Consonant t-SNE, seed 0, coloured on manner above and on place below

A.5 Complete Models BLEU Score Tables

The tables introduced here are the complete BLEU score tables for all our models language pairs, in 1-best and 10-best prediction. The standard deviation and mean are computed across all data shufflings used to train our models. These tables therefore represent 255 models (81 language directions * 3 bilingual models * 3 shuffling seeds, + 4 multilingual models trained on all directions at once * 3 shuffling seeds).

FROM CA TO	CA	ES	FR	GL	IT	OC	PT	RO	RUP
1-best									
SMT	100.0 ± 0.0	72.0 ± 3.6	68.4 ± 2.3	63.4 ± 0.8	57.3 ± 0.6	85.0 ± 5.8	74.2 ± 3.0	32.6 ± 10.7	39.4 ± 3.7
Bi-NMT	99.6 ± 0.1	64.1 ± 3.4	45.0 ± 4.8	34.7 ± 2.3	43.9 ± 3.0	39.2 ± 7.8	52.8 ± 1.5	5.7 ± 3.0	4.8 ± 0.3
Bi-NMT+m	99.6 ± 0.1	74.0 ± 1.5	60.7 ± 4.6	58.4 ± 2.8	53.4 ± 2.7	77.6 ± 9.9	73.9 ± 2.9	19.9 ± 15.6	19.7 ± 8.2
M-NMT	*nan* ± *nan*	64.9 ± 2.7	61.2 ± 5.9	58.7 ± 4.1	52.7 ± 1.7	63.2 ± 2.0	63.3 ± 4.5	38.4 ± 1.2	46.9 ± 5.5
M-NMT+m	89.7 ± 0.9	74.6 ± 2.4	74.5 ± 4.3	73.0 ± 2.5	58.8 ± 0.4	75.9 ± 4.3	77.2 ± 2.4	50.2 ± 11.4	49.2 ± 8.0
+shared_emb	89.2 ± 1.7	74.0 ± 0.1	73.4 ± 1.8	67.0 ± 2.7	62.1 ± 0.9	84.9 ± 5.7	77.0 ± 5.0	39.3 ± 11.6	47.3 ± 7.7
+shared_all	59.3 ± 1.3	65.0 ± 2.8	66.7 ± 4.9	62.4 ± 4.3	51.5 ± 1.7	81.2 ± 5.8	69.2 ± 3.6	45.0 ± 11.7	43.7 ± 6.5
10-best									
SMT	100.0 ± 0.0	89.8 ± 0.6	86.6 ± 2.5	81.2 ± 3.2	81.3 ± 2.3	90.2 ± 4.7	91.4 ± 2.1	63.7 ± 10.4	57.2 ± 4.5
Bi-NMT	99.9 ± 0.1	85.4 ± 1.2	69.9 ± 3.9	56.1 ± 3.9	64.1 ± 3.8	63.5 ± 5.3	78.3 ± 1.2	20.4 ± 9.6	12.6 ± 3.1
Bi-NMT+m	99.9 ± 0.1	90.2 ± 0.3	81.0 ± 2.2	76.4 ± 2.8	76.5 ± 2.7	83.8 ± 8.6	88.6 ± 2.2	35.4 ± 18.8	35.9 ± 2.9
M-NMT	*nan* ± *nan*	87.1 ± 1.4	84.3 ± 3.6	79.6 ± 3.4	77.2 ± 1.6	80.6 ± 1.2	86.1 ± 3.1	63.0 ± 6.0	71.8 ± 3.6
M-NMT+m	97.8 ± 0.3	90.9 ± 1.5	89.9 ± 3.1	89.8 ± 3.1	85.7 ± 1.3	88.8 ± 4.1	92.4 ± 1.1	71.2 ± 7.9	74.5 ± 6.9
+shared_emb	97.9 ± 0.5	91.8 ± 0.9	89.6 ± 3.4	84.4 ± 3.8	88.0 ± 0.4	92.5 ± 4.5	91.7 ± 1.3	66.1 ± 6.4	77.1 ± 6.9
+shared_all	75.2 ± 1.0	87.6 ± 1.3	89.7 ± 2.0	83.9 ± 4.1	77.1 ± 2.0	93.9 ± 3.4	89.9 ± 1.6	63.8 ± 8.4	73.5 ± 10.9
FROM ES TO	CA	ES	FR	GL	IT	OC	PT	RO	RUP
1-best									
SMT	71.2 ± 0.4	100.0 ± 0.0	62.4 ± 0.9	67.4 ± 4.1	63.0 ± 0.5	48.6 ± 9.4	76.7 ± 2.6	34.4 ± 3.4	38.3 ± 5.5
Bi-NMT	73.9 ± 4.6	99.5 ± 0.1	51.6 ± 3.4	56.0 ± 3.0	57.7 ± 2.6	3.0 ± 0.2	65.9 ± 8.3	19.2 ± 5.1	5.7 ± 2.6
Bi-NMT+m	81.2 ± 2.9	99.5 ± 0.1	59.1 ± 4.4	69.4 ± 0.8	67.2 ± 2.2	37.8 ± 3.4	76.7 ± 2.6	26.2 ± 1.0	22.9 ± 13.1
M-NMT	72.1 ± 4.7	*nan* ± *nan*	57.5 ± 2.7	70.5 ± 4.4	53.4 ± 2.5	75.7 ± 9.5	69.0 ± 3.7	37.6 ± 7.7	48.8 ± 9.0
M-NMT+m	79.0 ± 1.9	88.6 ± 1.1	67.3 ± 2.0	72.1 ± 6.2	63.1 ± 1.2	86.1 ± 3.3	73.7 ± 2.1	46.8 ± 2.7	45.9 ± 6.4
+shared_emb	80.8 ± 0.8	90.3 ± 2.5	71.4 ± 0.2	74.8 ± 2.6	64.8 ± 1.2	84.2 ± 8.0	76.4 ± 4.8	48.2 ± 5.6	42.4 ± 8.4
+shared_all	72.4 ± 3.0	61.8 ± 0.4	64.5 ± 1.9	67.3 ± 3.5	49.5 ± 3.7	78.7 ± 8.9	69.8 ± 2.5	38.2 ± 5.2	42.2 ± 8.1
10-best									
SMT	90.3 ± 1.6	100.0 ± 0.0	79.6 ± 2.5	87.2 ± 2.1	86.3 ± 0.8	78.0 ± 5.4	91.9 ± 0.9	60.4 ± 6.8	53.7 ± 7.1
Bi-NMT	89.3 ± 2.8	100.0 ± 0.0	69.6 ± 2.3	75.8 ± 1.1	82.7 ± 2.4	8.8 ± 1.8	85.2 ± 6.2	44.1 ± 0.8	14.0 ± 5.9
Bi-NMT+m	91.9 ± 1.9	100.0 ± 0.0	79.0 ± 1.0	84.7 ± 2.3	86.4 ± 2.3	60.0 ± 2.6	91.4 ± 1.1	48.3 ± 2.4	41.6 ± 8.2
M-NMT	89.9 ± 2.5	*nan* ± *nan*	80.6 ± 4.2	86.5 ± 4.6	80.0 ± 2.0	92.5 ± 4.5	87.6 ± 2.0	62.4 ± 7.2	71.0 ± 6.7
M-NMT+m	93.8 ± 1.4	97.9 ± 0.4	83.8 ± 2.0	88.8 ± 3.3	86.1 ± 0.1	94.9 ± 2.5	91.5 ± 0.7	68.4 ± 6.9	69.2 ± 3.1
+shared_emb	93.9 ± 1.1	98.6 ± 0.5	85.5 ± 2.8	90.6 ± 3.1	87.2 ± 0.6	91.8 ± 6.8	93.5 ± 2.6	71.0 ± 2.9	69.3 ± 5.6
+shared_all	91.4 ± 2.0	79.9 ± 1.5	80.2 ± 4.0	88.6 ± 4.2	80.0 ± 1.8	92.6 ± 4.7	91.2 ± 0.6	64.5 ± 2.6	65.9 ± 4.5
FROM FR TO	CA	ES	FR	GL	IT	OC	PT	RO	RUP
1-best									
SMT	67.7 ± 2.7	63.4 ± 1.1	100.0 ± 0.0	55.9 ± 6.7	50.0 ± 3.9	32.6 ± 5.3	58.4 ± 2.9	21.5 ± 2.3	18.5 ± 6.8
Bi-NMT	40.1 ± 3.6	39.3 ± 5.4	98.7 ± 0.4	10.0 ± 5.8	28.9 ± 3.4	5.1 ± 0.7	31.2 ± 7.5	3.8 ± 1.5	2.3 ± 0.3
Bi-NMT+m	62.1 ± 3.2	58.1 ± 5.9	98.7 ± 0.4	34.3 ± 4.4	48.1 ± 5.4	7.2 ± 2.6	51.0 ± 2.1	8.4 ± 2.3	8.8 ± 2.9
M-NMT	66.0 ± 3.8	53.7 ± 2.6	*nan* ± *nan*	62.8 ± 6.9	45.6 ± 3.2	62.8 ± 8.3	54.8 ± 3.5	21.8 ± 6.4	30.9 ± 19.8
M-NMT+m	74.9 ± 7.9	64.5 ± 1.5	83.8 ± 1.6	68.7 ± 4.9	53.2 ± 4.3	75.9 ± 10.8	64.8 ± 2.1	28.4 ± 3.0	21.4 ± 13.3
+shared_emb	70.9 ± 3.8	65.9 ± 4.1	81.9 ± 4.3	69.5 ± 5.6	56.3 ± 3.9	81.3 ± 10.3	65.2 ± 3.0	34.6 ± 6.4	14.5 ± 5.6
+shared_all	66.3 ± 3.8	54.0 ± 4.0	53.0 ± 5.7	57.9 ± 4.4	46.1 ± 5.4	67.3 ± 5.6	54.6 ± 2.0	28.0 ± 9.5	18.4 ± 8.8
10-best									
SMT	85.1 ± 0.9	79.9 ± 3.1	100.0 ± 0.0	72.7 ± 5.5	70.9 ± 4.4	60.1 ± 2.8	77.1 ± 2.4	32.1 ± 10.9	28.4 ± 12.7
Bi-NMT	59.5 ± 1.7	60.5 ± 5.8	99.2 ± 0.3	24.7 ± 5.6	49.9 ± 7.4	9.2 ± 1.2	51.4 ± 7.8	8.6 ± 1.3	9.1 ± 0.8
Bi-NMT+m	79.0 ± 2.6	73.2 ± 5.7	99.2 ± 0.3	55.5 ± 5.0	66.7 ± 6.1	21.1 ± 5.1	69.4 ± 1.1	15.5 ± 5.8	23.6 ± 18.4
M-NMT	83.6 ± 2.3	79.8 ± 2.0	*nan* ± *nan*	82.2 ± 5.5	70.2 ± 4.4	81.4 ± 2.5	76.7 ± 2.6	46.6 ± 8.7	60.4 ± 27.2
M-NMT+m	89.8 ± 4.0	85.8 ± 1.9	94.9 ± 0.9	86.7 ± 3.0	78.8 ± 1.1	85.1 ± 12.1	82.0 ± 2.8	57.8 ± 2.5	54.4 ± 20.8
+shared_emb	89.4 ± 2.1	84.9 ± 2.3	93.3 ± 2.0	88.2 ± 5.6	76.9 ± 3.7	95.5 ± 3.2	80.7 ± 1.1	64.1 ± 7.9	42.1 ± 16.1
+shared_all	84.7 ± 3.8	76.7 ± 4.5	66.8 ± 4.2	82.2 ± 4.2	66.8 ± 3.3	92.5 ± 5.4	74.8 ± 0.7	47.0 ± 10.9	40.7 ± 14.3
FROM GL TO	CA	ES	FR	GL	IT	OC	PT	RO	RUP
1-best									
SMT	59.6 ± 4.1	74.9 ± 4.2	56.4 ± 9.0	100.0 ± 0.0	57.7 ± 6.6	54.6 ± 8.1	86.4 ± 1.6	29.7 ± 8.1	46.1 ± 13.6
Bi-NMT	38.8 ± 3.9	58.9 ± 3.0	11.4 ± 5.6	98.9 ± 1.3	30.5 ± 3.1	3.6 ± 0.5	72.7 ± 4.4	6.6 ± 1.0	4.7 ± 1.2
Bi-NMT+m	63.2 ± 1.7	73.2 ± 4.4	40.9 ± 7.2	98.9 ± 1.3	48.9 ± 7.7	22.6 ± 6.7	85.0 ± 0.7	15.2 ± 5.3	19.8 ± 2.2
M-NMT	69.0 ± 1.1	68.6 ± 4.7	59.6 ± 4.9	*nan* ± *nan*	56.3 ± 3.8	67.9 ± 9.2	75.5 ± 1.9	45.7 ± 13.6	39.6 ± 4.6
M-NMT+m	69.4 ± 3.5	72.8 ± 2.8	64.1 ± 8.6	86.6 ± 5.0	59.8 ± 6.8	71.3 ± 14.2	82.9 ± 1.9	52.1 ± 11.0	62.3 ± 6.2
+shared_emb	72.7 ± 2.1	74.3 ± 1.5	61.5 ± 11.3	91.1 ± 1.1	62.6 ± 0.9	75.5 ± 4.0	87.1 ± 0.6	57.5 ± 11.6	57.1 ± 17.6
+shared_all	68.3 ± 3.5	68.9 ± 3.9	55.9 ± 10.7	64.2 ± 5.7	59.1 ± 5.8	69.3 ± 10.9	78.7 ± 3.9	51.0 ± 11.3	59.5 ± 4.7
10-best									
SMT	85.8 ± 0.4	89.0 ± 1.6	72.5 ± 4.4	100.0 ± 0.0	77.0 ± 5.7	78.1 ± 6.8	93.9 ± 2.2	59.3 ± 5.1	58.5 ± 14.5
Bi-NMT	58.9 ± 2.5	79.3 ± 2.4	22.8 ± 5.1	99.5 ± 0.6	48.3 ± 2.2	8.1 ± 2.7	87.2 ± 2.4	11.7 ± 2.3	12.6 ± 4.8
Bi-NMT+m	77.1 ± 1.1	87.5 ± 0.8	53.7 ± 8.3	99.5 ± 0.6	68.0 ± 6.6	48.0 ± 9.6	93.9 ± 1.0	31.1 ± 6.1	42.4 ± 7.9
M-NMT	85.3 ± 4.7	85.7 ± 2.7	76.3 ± 5.6	*nan* ± *nan*	79.6 ± 5.3	89.5 ± 9.2	93.4 ± 2.0	66.3 ± 4.2	81.3 ± 3.0
M-NMT+m	89.3 ± 3.7	89.5 ± 2.4	86.4 ± 5.3	96.4 ± 2.1	82.2 ± 5.1	88.2 ± 5.3	96.4 ± 2.3	77.0 ± 4.8	85.0 ± 6.8
+shared_emb	91.1 ± 1.0	90.2 ± 2.0	84.9 ± 2.8	98.7 ± 0.6	85.3 ± 5.2	94.1 ± 1.4	95.2 ± 1.1	78.8 ± 5.5	80.9 ± 5.6
+shared_all	88.2 ± 5.3	84.7 ± 1.1	80.4 ± 6.6	85.2 ± 4.2	76.9 ± 6.3	87.3 ± 7.8	93.3 ± 2.6	68.5 ± 7.0	80.0 ± 5.0
FROM IT TO	CA	ES	FR	GL	IT	OC	PT	RO	RUP
1-best									
SMT	63.3 ± 3.1	74.8 ± 1.7	61.6 ± 2.8	58.2 ± 7.5	100.0 ± 0.0	44.7 ± 13.8	70.4 ± 3.1	48.6 ± 3.1	49.2 ± 0.9
Bi-NMT	35.5 ± 3.9	70.8 ± 0.6	31.7 ± 8.6	30.7 ± 2.9	99.6 ± 0.1	6.5 ± 5.2	61.5 ± 1.3	29.8 ± 2.6	21.9 ± 5.0
Bi-NMT+m	68.0 ± 0.8	73.0 ± 2.8	59.6 ± 6.4	55.2 ± 7.8	99.6 ± 0.1	35.6 ± 14.6	70.6 ± 1.5	44.7 ± 4.3	34.1 ± 4.7
M-NMT	61.0 ± 4.3	60.0 ± 4.8	55.1 ± 3.8	61.6 ± 4.0	*nan* ± *nan*	55.8 ± 4.7	58.7 ± 3.5	51.9 ± 2.9	50.6 ± 3.8
M-NMT+m	73.3 ± 1.4	72.3 ± 1.7	64.3 ± 7.5	69.1 ± 5.4	81.8 ± 0.9	73.4 ± 5.8	72.9 ± 3.3	51.7 ± 2.2	52.8 ± 5.0
+shared_emb	72.8 ± 0.5	70.2 ± 3.9	66.5 ± 4.1	69.3 ± 5.4	81.4 ± 1.5	73.4 ± 8.3	73.5 ± 3.5	58.9 ± 2.8	50.9 ± 1.9
+shared_all	68.9 ± 4.1	60.8 ± 0.8	54.0 ± 6.1	59.6 ± 8.3	70.0 ± 3.3	71.2 ± 18.2	62.5 ± 2.1	44.2 ± 1.8	44.2 ± 1.1

Table 7: Results of our different models for the cognate prediction task - 1

FROM IT TO	CA	ES	FR	GL	IT	OC	PT	RO	RUP
10-best									
SMT	83.8 ± 2.1	89.1 ± 0.4	76.7 ± 3.0	78.3 ± 6.5	100.0 ± 0.0	68.1 ± 9.7	87.9 ± 1.7	70.2 ± 4.6	70.6 ± 1.5
Bi-NMT	56.0 ± 5.7	85.2 ± 1.6	53.4 ± 8.4	50.1 ± 1.5	99.9 ± 0.1	12.4 ± 4.2	83.5 ± 2.4	51.7 ± 1.7	41.2 ± 6.4
Bi-NMT+m	82.8 ± 0.9	87.3 ± 1.1	77.4 ± 6.4	74.8 ± 3.5	99.9 ± 0.1	51.1 ± 15.5	86.2 ± 0.6	67.2 ± 2.4	58.0 ± 3.8
M-NMT	81.8 ± 1.5	82.2 ± 3.0	76.5 ± 5.0	81.4 ± 4.4	*nan* ± *nan*	79.9 ± 2.9	81.9 ± 2.1	70.5 ± 4.5	72.7 ± 4.4
M-NMT+m	90.4 ± 1.8	88.0 ± 0.6	80.0 ± 3.4	86.6 ± 2.4	96.7 ± 0.8	84.1 ± 9.8	90.0 ± 0.9	80.1 ± 1.4	73.4 ± 1.0
+shared_emb	89.6 ± 0.6	89.4 ± 1.7	80.6 ± 3.9	87.3 ± 3.1	96.5 ± 0.6	85.7 ± 9.3	89.7 ± 1.5	77.1 ± 2.0	72.2 ± 0.3
+shared_all	83.5 ± 0.7	81.3 ± 2.0	76.6 ± 7.1	80.6 ± 4.5	91.9 ± 1.6	83.3 ± 10.1	87.3 ± 1.4	71.4 ± 4.0	67.4 ± 6.9
FROM OC TO	CA	ES	FR	GL	IT	OC	PT	RO	RUP
1-best									
SMT	88.2 ± 1.8	57.8 ± 7.1	34.1 ± 5.0	57.5 ± 9.3	53.1 ± 3.0	100.0 ± 0.0	44.0 ± 6.0	21.2 ± 10.4	30.7 ± 13.8
Bi-NMT	60.6 ± 10.6	7.3 ± 1.1	3.4 ± 1.4	4.1 ± 2.0	8.2 ± 2.6	97.8 ± 1.1	4.0 ± 0.9	3.2 ± 1.4	4.6 ± 1.4
Bi-NMT+m	84.9 ± 1.2	42.4 ± 4.7	11.6 ± 6.1	19.1 ± 6.2	42.9 ± 2.5	97.8 ± 1.1	39.5 ± 7.4	10.2 ± 2.4	7.5 ± 0.3
M-NMT	75.2 ± 8.8	56.7 ± 7.8	49.1 ± 11.0	64.7 ± 8.0	55.4 ± 2.1	*nan* ± *nan*	59.4 ± 2.6	47.3 ± 6.5	69.9 ± 5.5
M-NMT+m	84.8 ± 2.4	69.5 ± 4.8	54.6 ± 5.5	71.5 ± 7.4	72.0 ± 4.5	82.3 ± 6.3	59.5 ± 10.6	58.9 ± 5.6	61.1 ± 5.0
+shared_emb	86.3 ± 7.1	73.8 ± 11.2	53.5 ± 1.5	76.1 ± 13.2	69.0 ± 7.1	84.2 ± 3.8	60.0 ± 16.2	70.1 ± 13.0	74.1 ± 5.3
+shared_all	86.5 ± 2.2	60.5 ± 10.0	41.2 ± 8.7	64.7 ± 10.3	58.4 ± 6.8	59.1 ± 3.4	57.2 ± 7.8	51.3 ± 18.7	57.5 ± 11.5
10-best									
SMT	92.4 ± 2.6	80.0 ± 8.4	42.2 ± 5.3	74.0 ± 8.2	71.5 ± 2.6	100.0 ± 0.0	72.1 ± 3.4	35.9 ± 10.4	45.8 ± 6.4
Bi-NMT	75.2 ± 6.1	13.6 ± 3.2	8.3 ± 4.6	7.7 ± 3.5	18.6 ± 3.3	99.4 ± 0.8	8.0 ± 1.6	8.4 ± 1.9	10.4 ± 1.3
Bi-NMT+m	93.0 ± 2.4	63.6 ± 8.3	19.5 ± 9.8	38.0 ± 17.4	61.3 ± 1.9	99.4 ± 0.8	53.4 ± 8.5	25.1 ± 8.4	17.4 ± 4.9
M-NMT	91.0 ± 6.5	85.3 ± 6.0	61.9 ± 9.1	79.7 ± 5.5	79.5 ± 2.5	*nan* ± *nan*	84.3 ± 3.9	76.4 ± 4.5	88.9 ± 11.5
M-NMT+m	94.9 ± 2.5	89.2 ± 6.0	70.5 ± 5.9	88.8 ± 6.4	88.5 ± 3.3	92.4 ± 3.1	86.7 ± 3.3	70.7 ± 4.2	88.1 ± 4.9
+shared_emb	97.1 ± 2.1	86.1 ± 7.2	67.9 ± 4.6	91.4 ± 3.2	85.6 ± 8.8	94.1 ± 1.3	86.8 ± 8.3	79.3 ± 4.0	86.0 ± 10.4
+shared_all	94.4 ± 2.4	83.1 ± 6.6	66.2 ± 5.1	85.1 ± 6.2	77.1 ± 6.2	72.0 ± 2.1	85.3 ± 3.1	71.5 ± 10.3	80.5 ± 12.3
FROM PT TO	CA	ES	FR	GL	IT	OC	PT	RO	RUP
1-best									
SMT	75.0 ± 0.1	75.4 ± 0.3	63.2 ± 5.0	89.2 ± 0.7	59.4 ± 5.9	50.8 ± 4.7	100.0 ± 0.0	42.2 ± 1.9	45.5 ± 2.3
Bi-NMT	66.0 ± 4.1	69.2 ± 1.0	39.0 ± 7.8	75.3 ± 3.5	50.8 ± 3.1	6.3 ± 1.6	99.3 ± 0.4	11.9 ± 5.7	10.9 ± 3.0
Bi-NMT+m	75.9 ± 3.0	74.9 ± 2.1	56.2 ± 2.7	86.0 ± 2.1	59.5 ± 4.2	29.2 ± 5.9	99.3 ± 0.4	28.8 ± 6.8	27.3 ± 3.8
M-NMT	74.0 ± 3.3	69.2 ± 2.3	63.9 ± 3.6	77.2 ± 0.3	55.4 ± 3.7	72.4 ± 6.6	*nan* ± *nan*	48.8 ± 6.4	62.1 ± 5.6
M-NMT+m	78.7 ± 3.9	75.8 ± 4.0	67.8 ± 0.5	83.9 ± 1.7	63.8 ± 1.6	89.1 ± 3.3	89.0 ± 1.7	55.7 ± 5.9	61.0 ± 12.6
+shared_emb	78.0 ± 3.4	73.1 ± 2.9	70.3 ± 4.1	82.2 ± 3.0	61.4 ± 2.3	81.9 ± 5.7	88.4 ± 1.9	52.9 ± 7.2	61.7 ± 3.2
+shared_all	76.4 ± 3.0	67.3 ± 0.7	63.4 ± 3.6	78.0 ± 3.7	55.1 ± 2.9	71.2 ± 5.4	64.2 ± 2.2	47.7 ± 5.6	56.1 ± 8.8
10-best									
SMT	86.9 ± 1.1	91.6 ± 0.7	83.1 ± 4.9	96.2 ± 1.0	80.9 ± 3.6	76.4 ± 9.6	100.0 ± 0.0	67.8 ± 4.8	74.2 ± 2.1
Bi-NMT	80.1 ± 3.3	88.5 ± 0.5	61.0 ± 5.2	89.1 ± 2.3	73.6 ± 2.5	11.7 ± 1.5	99.8 ± 0.1	24.2 ± 1.3	36.5 ± 3.3
Bi-NMT+m	86.5 ± 2.7	89.5 ± 0.8	76.0 ± 4.0	93.9 ± 1.6	82.0 ± 3.6	43.6 ± 3.7	99.8 ± 0.1	43.2 ± 4.6	51.3 ± 2.4
M-NMT	88.5 ± 2.2	89.0 ± 1.4	85.8 ± 2.8	93.0 ± 1.1	80.0 ± 3.6	90.4 ± 2.4	*nan* ± *nan*	70.3 ± 5.9	83.8 ± 2.2
M-NMT+m	90.0 ± 3.1	92.1 ± 1.0	86.6 ± 3.0	94.5 ± 1.9	85.1 ± 2.1	96.4 ± 4.3	98.7 ± 0.7	77.8 ± 4.2	80.3 ± 11.6
+shared_emb	89.7 ± 2.8	91.4 ± 1.0	89.0 ± 2.6	95.8 ± 1.3	85.2 ± 2.8	95.4 ± 3.9	97.7 ± 1.1	73.6 ± 9.8	84.4 ± 3.2
+shared_all	87.0 ± 1.1	88.6 ± 2.3	85.5 ± 1.9	92.9 ± 1.3	75.9 ± 2.1	93.1 ± 4.2	84.6 ± 3.0	69.6 ± 2.8	85.0 ± 1.5
FROM RO TO	CA	ES	FR	GL	IT	OC	PT	RO	RUP
1-best									
SMT	32.9 ± 5.3	37.6 ± 5.8	20.2 ± 2.6	29.7 ± 10.4	43.5 ± 5.6	25.5 ± 2.8	32.7 ± 0.6	100.0 ± 0.0	66.3 ± 1.7
Bi-NMT	10.4 ± 3.4	22.6 ± 4.5	6.3 ± 1.4	2.1 ± 0.5	33.1 ± 8.7	7.1 ± 2.5	14.9 ± 6.2	98.5 ± 1.4	59.0 ± 8.3
Bi-NMT+m	18.1 ± 4.8	34.2 ± 3.0	7.2 ± 3.3	15.9 ± 2.5	44.7 ± 7.0	12.9 ± 1.8	21.7 ± 7.3	98.5 ± 1.4	67.4 ± 9.8
M-NMT	47.9 ± 2.4	48.5 ± 2.1	37.9 ± 9.2	47.0 ± 3.9	42.0 ± 6.6	51.0 ± 17.3	42.0 ± 5.6	*nan* ± *nan*	58.1 ± 8.3
M-NMT+m	47.2 ± 4.0	56.4 ± 7.4	36.9 ± 10.7	55.6 ± 4.1	53.2 ± 2.2	59.1 ± 13.5	45.7 ± 3.4	70.4 ± 2.3	70.7 ± 9.4
+shared_emb	57.7 ± 7.1	54.2 ± 3.7	36.0 ± 4.7	54.6 ± 6.6	55.1 ± 4.9	63.0 ± 13.4	50.7 ± 6.3	70.4 ± 1.8	75.6 ± 8.0
+shared_all	53.7 ± 5.3	33.1 ± 5.6	37.8 ± 6.3	50.9 ± 6.8	37.3 ± 2.2	56.8 ± 11.5	38.4 ± 7.3	48.1 ± 0.9	63.4 ± 7.4
10-best									
SMT	57.9 ± 3.9	63.7 ± 7.6	38.1 ± 6.1	47.0 ± 6.4	72.1 ± 4.2	44.5 ± 9.6	58.3 ± 2.3	100.0 ± 0.0	87.4 ± 2.0
Bi-NMT	22.5 ± 10.8	45.4 ± 0.7	10.0 ± 0.4	6.0 ± 0.2	58.1 ± 5.6	14.2 ± 3.8	30.8 ± 4.8	99.6 ± 0.5	80.8 ± 9.8
Bi-NMT+m	38.2 ± 8.4	58.3 ± 4.5	16.2 ± 5.2	32.9 ± 8.8	64.9 ± 4.3	27.2 ± 3.9	51.5 ± 4.0	99.6 ± 0.5	85.7 ± 8.8
M-NMT	79.6 ± 4.8	75.7 ± 5.9	56.6 ± 16.0	66.9 ± 2.0	71.3 ± 4.5	74.7 ± 15.3	70.2 ± 3.7	*nan* ± *nan*	80.1 ± 9.2
M-NMT+m	75.9 ± 5.6	80.5 ± 8.0	52.8 ± 8.8	76.2 ± 5.9	80.8 ± 3.9	77.9 ± 7.5	75.8 ± 3.7	89.3 ± 3.3	87.2 ± 4.9
+shared_emb	80.8 ± 5.5	82.7 ± 4.6	65.2 ± 6.1	81.0 ± 5.3	82.4 ± 2.1	83.0 ± 14.0	76.0 ± 2.4	89.5 ± 1.0	90.2 ± 7.0
+shared_all	74.6 ± 9.8	64.5 ± 6.2	60.8 ± 7.1	69.7 ± 8.4	67.0 ± 3.3	66.9 ± 14.3	68.3 ± 4.6	64.5 ± 1.6	84.8 ± 6.3
FROM RUP TO	CA	ES	FR	GL	IT	OC	PT	RO	RUP
1-best									
SMT	29.2 ± 2.4	32.4 ± 1.9	21.7 ± 2.9	29.5 ± 13.2	36.6 ± 4.1	26.1 ± 12.4	42.0 ± 5.5	63.3 ± 7.3	100.0 ± 0.0
Bi-NMT	2.7 ± 0.7	3.3 ± 0.7	5.7 ± 1.0	3.1 ± 1.9	26.7 ± 2.6	5.2 ± 2.0	27.1 ± 3.0	48.8 ± 5.1	95.2 ± 1.8
Bi-NMT+m	16.4 ± 4.5	23.4 ± 1.9	9.1 ± 1.7	15.4 ± 9.1	30.6 ± 0.4	14.4 ± 5.3	28.9 ± 12.6	64.8 ± 5.4	95.2 ± 1.8
M-NMT	50.1 ± 12.7	36.7 ± 6.3	32.0 ± 12.7	33.4 ± 1.9	44.4 ± 4.9	29.9 ± 3.1	56.8 ± 5.6	57.7 ± 3.0	*nan* ± *nan*
M-NMT+m	60.0 ± 4.8	51.8 ± 7.4	24.6 ± 14.4	49.6 ± 8.0	44.7 ± 3.5	63.5 ± 7.9	60.4 ± 7.1	67.9 ± 4.7	70.4 ± 6.0
+shared_emb	59.2 ± 8.4	47.2 ± 3.5	46.7 ± 5.0	54.6 ± 6.7	48.9 ± 4.3	41.7 ± 11.4	61.6 ± 5.6	66.7 ± 3.4	75.6 ± 3.2
+shared_all	46.9 ± 20.6	25.1 ± 6.7	35.2 ± 18.3	37.3 ± 12.1	34.0 ± 5.0	53.6 ± 9.9	39.0 ± 12.7	52.6 ± 6.4	59.8 ± 2.1
10-best									
SMT	53.8 ± 14.2	60.4 ± 7.7	32.4 ± 11.5	45.7 ± 6.8	62.6 ± 0.7	35.2 ± 11.3	62.7 ± 9.1	83.1 ± 7.4	100.0 ± 0.0
Bi-NMT	8.3 ± 4.1	15.6 ± 6.8	13.4 ± 3.8	7.0 ± 2.3	44.6 ± 2.0	7.3 ± 1.0	46.6 ± 5.3	72.0 ± 7.0	98.4 ± 1.3
Bi-NMT+m	25.1 ± 6.0	51.8 ± 4.7	17.8 ± 5.4	22.1 ± 10.9	51.9 ± 1.8	31.1 ± 15.8	51.8 ± 9.9	80.9 ± 8.1	98.4 ± 1.3
M-NMT	77.4 ± 9.0	72.3 ± 1.5	62.2 ± 11.2	66.9 ± 8.0	69.4 ± 6.0	46.7 ± 12.2	79.0 ± 1.5	79.5 ± 0.7	*nan* ± *nan*
M-NMT+m	73.6 ± 10.6	80.1 ± 6.5	53.4 ± 18.4	78.7 ± 12.1	72.5 ± 3.3	77.2 ± 7.1	81.6 ± 5.6	83.2 ± 4.0	89.2 ± 4.4
+shared_emb	79.2 ± 12.5	78.6 ± 11.0	63.4 ± 9.6	77.6 ± 7.2	74.7 ± 1.9	82.3 ± 3.3	80.8 ± 1.5	83.4 ± 5.6	89.9 ± 3.2
+shared_all	69.1 ± 13.9	60.9 ± 7.2	62.4 ± 14.9	62.3 ± 1.5	64.0 ± 3.4	73.6 ± 18.8	72.9 ± 4.8	77.9 ± 8.4	76.4 ± 0.9

Table 8: Results of our different models for the cognate prediction task - 2