Exploring the Low-Resource Transfer-Learning with mT5 model

Anonymous ACL submission

Abstract

Languages are mortal. While the NLP community tends to expand its competence to multilingual models, there is still a great risk for low-resource languages to vanish before any prototypes appear for them.

This paper presents a series of experiments that explore the transfer learning for low-resource languages, testing hypotheses about finding the optimal donor language on the typological relations and grammatical features. Our results showed that multilingual models like mT5 obtain significantly lower perplexity on 45/46 low-resource languages without training on them.

We collected the most variable multilingual training corpus available with 288 languages, based on the linguistically-wise databases, field linguist resources, the World Atlas of Language Structures, and Wikipedia.

1 Introduction

Multilingualism has become a major trend in NLP in general and language modelling in particular. The inestimable value of computational multilingualism lies both in the striving for the ability to generalize on any natural language possible and to explore the civilizational and intellectual heritage of the preserved cultures.

The preservation of linguistic diversity depends on a large number of factors, including economic and political ones, and is a subject of discussion[1]. However, in this work, we focus on the possibilities within language technologies for sustaining language use.

According to the Endangered Languages Project (Belew, 2021), more than 3000 languages are in the state of risk at least. The efforts of the NLP community in recent years have undoubtedly expanded the toolkit for working with languages: on the one hand, large transformer models such as mBert (Devlin et al., 2018) and mT5 (Xue et al., 2020) have significantly expanded the possibilities of working with languages using transfer learning and benchmarks that require models to be cross-lingual (XGLUE (Liang et al., 2020b,a), XTREME (Ruder et al., 2021)) require such models to be able to solve tasks for the entire typologically weighted sample of languages. However, the inequality of the language use in the NLP research can be described by the power law relation between language frequency and its rank, but with longer tail, see Figure 1).

Our data crawling showed that there is a large number of "twilight zone" languages that have a non-zero representation in the data catalogs, corpora, or local websites and yet have not been ever included in the language modelling research. The list of these languages, including the low-resource ones, can be found in this paper. We believe that the introduction of these languages into the general practice of language technology can help equalize existing biases, bring them to the attention of the international community and stimulate the creation of new tools and new data available in these languages.

We denote the **donor** language as a source for fine-tuning the language model, and the **target** language as a material for testing the model after the tuning. The donor languages in this work were chosen from high-resource (HR) languages whose textual data will be used for the model training during the transfer-learning experiments. Target languages are low-resource (LR).

We thus postulate that the main contributions of this work are as follows:

- a series of hypotheses testing on the choice of the best available donor languages for the low-resource languages.
- the methodology for collecting the most mul-

[1] Towards a New Language of the Global Language Crisis: shorturl.at/cgjuC

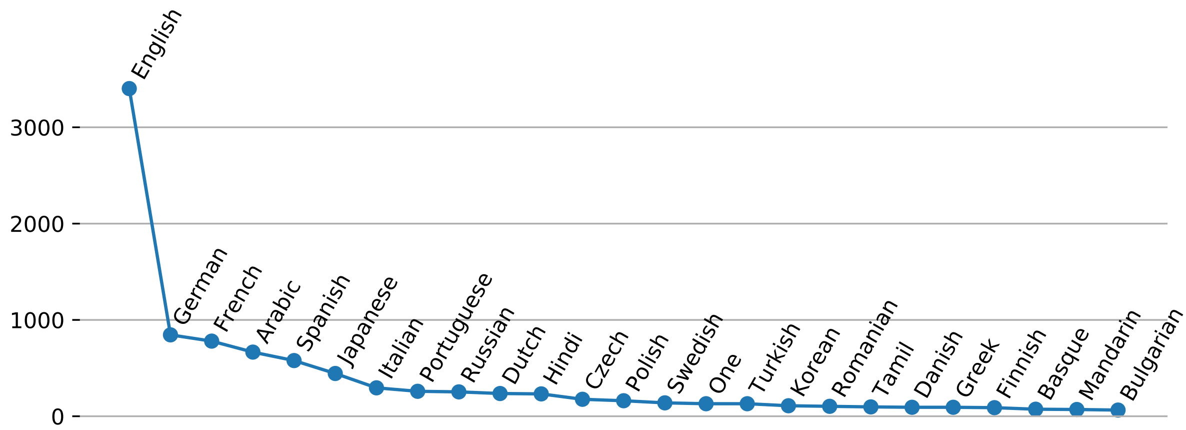


Figure 1: Language citation inequality in ACL Anthology papers, based on the language list from WALS.info (top 25, from 1989 and older to Oct 2021)

tilingual corpus of the existing, used for teaching and/or testing language models. The corpus is assembled based on existing data from various projects;

- an independent analysis of mT5 performance and its limits in low-resource language modelling task.

2 Previous Work

The expansion of the NLP methods to other languages leans heavily on 1) transfer learning techniques and 2) multilingual data standardization. While the data will be discussed in Section 3, we will now dwell on previous works that raise the issue of LR language modelling.

While there are 104 languages in mBert (Devlin et al., 2018) and 101 languages in mT5 (Xue et al., 2020), the models' ability to transfer knowledge and skills is criticized. (Libovický et al., 2019) shows that mBert context embeddings are not directly usable for zero-shot cross-lingual tasks for the languages in the training data. The work (Wu and Dredze, 2020) especially focuses on the quality of representation for LR languages, measured by within-language performance. It is stated that mBert doesn't perform for them as well as for HR languages, and the languages are not represented equally. Nevertheless, the situation can be improved by pairing the low-resource languages with their closest HR relative, for example, Lithuanian with Latvian and Dutch with Afrikaans.

Compared to mBert, several more modern language models were scored in the setting of typologically motivated transfer-learning. (Lauscher et al., 2020) show that typological motivation in choice of languages positively impacts the transfer learning scores. (Lin et al., 2019) provide a tool to choose such languages, which is capable of the direct usage of the languages typological characteristics. (Turc et al., 2021) further demonstrate that the languages beyond English, especially the resourced ones, are capable of reliable transfer learning, and could even outperform English as a source language in some cases. Unfortunately the direct study of the endangered languages behavior, which compose the long tail of the popular languages list, were out of the scope of both studies.

And even the promising amount of languages can still be considered a typologically biased subset from the overall variety: at least 2600+ languages described in the World Atlas of Language Structures database (Dryer and Haspelmath, 2013), and 7000+ in Ethnologue (Eberhard et al., 2021). For these languages, there at least do exist grammar descriptions and example sentences, texts, or speech audio.

As developing pre-trained language models for LR languages remains an open challenge, we consider **the search of the most proper HR donor language for those target LR languages available**. The major hypothesis is to take the closest relative as a donor language. But this only works for some families as there are a lot of families consisting of LR languages only.

Works that explore other strategies appear recently: (Kocmi and Bojar, 2018) proposes the usage vocabulary overlaps for finding a better HR donor. The results show that the method provides improvements for the machine-translation task even for totally unrelated language pairs, although the improvement is not always significant.

Based on the results of the aforementioned works, we propose the following hypotheses for testing:

- training big language models on the donor language can lead to better generalization on the target language data, measured with perplexity;
- donor languages should be HR and taken from the same genus for lowering perplexity on the target LR languages;
- if there are no HR donor candidates from the same genus, typological alternative can be used: e.g. a HR language with the most similar grammatical features.

In our series of experiments, we focused specifically on the task of language modelling: in the absence of labelled data for more applied tasks for all the LR languages, we considered the task of language modelling to be the major one for testing the suitability of our approach.

The data and the motivation on the low-resource language choice are described in Section 3. The experimental setup of the hypothesis testing is described in Section 4.

3 Multilingual Data

As part of the project, we decided to use a wide range of linguistic resources, in addition to commonly used corpora. We deliberately did not include such projects as Oscar (Ortiz Suárez et al., 2019) and Cleaned Colossal Common Crawl (Raffel et al., 2019) in the sources since, on the one hand, they are already partially represented in the training set of large language models. On the other hand, their language classification is obtained automatically. Such an approach can have an extremely harmful effect on LR languages that are not included in the class list but are nevertheless present in the sample.

The general corpus includes text materials of the following projects:

1. Wikipedia in every language available (CC BY-SA);
2. Universal Dependencies project[2] (de Marneffe et al., 2021) (original texts without annotation, the license for every treebank is different, mainly GNU GPL 3.0/LGPLLR/CC BY-based);
3. The Hamburg Center for Language Corpora (HZSK-PUB)[3] (linguistic primary research textual data, not restricted by copyright or personal data protection);
4. The Endangered Languages Archive[4] (text content only, no multimedia, non-commercial private research or educational activity);
5. Collected and annotated corpora of the languages of CIB countries [5] (Krylova et al., 2015).

And the final list of collected languages reaches a value of 288 languages with available textual data, which are listed in Appendix 1. We also measured the number of symbols and symbols in textual data for each collected language using a tokenizer of mT5-Base (Xue et al., 2020) model. The language statistics are available in Appendix 2.

3.1 Choice of Low-Resource Languages

While the LR language programs are based on a range of socioeconomic criteria for prioritizing choice of one or another language for support (Cieri et al., 2016), the computational approach to low-resource definition is mostly quantitatively oriented. As (Hedderich et al., 2020) states, different data thresholds are used to define LR: for various downstream tasks, like part-of-speech tagging, the limiting principle is the annotators' time - in (Garrette and Baldridge, 2013) it is restricted to 2 hours resulting in up to 1-2k tokens. For weakly supervised learning on the same tasks, the definition of the LR is updated to 10k labelled tokens (Kann et al., 2020). For unsupervised learning, for example, training language models for generation tasks, this limit is more exacting: 350k tokens (Yang et al., 2019).

To form the final list of LR languages, we decided to fix the LR languages' boundaries as 10k tokens for the lower bound and 350k tokens as the upper bound. According to that rules, the final list of LR languages includes 46 samples: *'Akan', 'Atikamekw', 'Bambara',*

[2] https://universaldependencies.org/

[3] https://corpora.uni-hamburg.de/hzsk/

[4] https://www.elararchive.org/

[5] Commonwealth of Independent States, http://web-corpora.net/

'Bhojpuri', 'Bislama', 'Cantonese', 'Chamorro', 'Cherokee', 'Cheyenne', 'Chichewa', 'Coptic', 'Dagbani', 'Ewe', 'Greenlandic (South)', Guaraní', 'Kashmiri', 'Kikuyu', 'Komi-Zyrian', 'Kongo', 'Koryak', 'Kurmanji', 'Madurese', 'Nadroga', 'Nanai', 'Nauruan', 'Quiché', 'Romani (Lovari)', 'Rundi', 'Samoan', 'Sango', 'Sesotho', 'Shor', 'Sranan', 'Swati', 'Tabassaran', 'Tahitian', 'Tat (Muslim)', 'Tigrinya', 'Tofa', 'Tok Pisin', 'Tsakhur', 'Tsonga', 'Udi', 'Venda', 'Yukaghir (Kolyma)', 'Zhuang (Northern)'.

4 Experiments

We conducted the experiments with the original 580M parameter mT5-Base model (Xue et al., 2020).

The main goal of our experiments was to figure out will the training on a high-resource donor help to better model a low-resource target or not. We formulated two different scenarios for measuring how the mT5 understands in the masked-language modelling (MLM) task:

—*Evaluation of the results of the original mt5-Base model on a LR language.* Here, we observe how well the model modulates a new language as is, just in a zero-shot manner.

—*Additional training of the model on a donor language in MLM task with further evaluation of the results on a LR language.* In this case, we want to oversee whether the training on the donor language will help the model to understand the LR language better or not. Of course, both languages must be somehow connected for achieving the best and logical result.

The idea behind our choice of the MLM as the transfer learning task is that all we have is the unlabeled raw textual data. And since our goal is to train the model to understand languages better, the MLM is the best match for it according to our setup. Also, the considered mT5 model was originally trained on MLM task.

4.1 Donor selection hypotheses

For each LR language, we need to choose the HR donors in order to follow the second scenario. We propose two different approaches of forming candidates (donors) for transfer learning:

- **Genetic proximity**. The donor is chosen from the same genus, as the LR language, based on the WALS[6] genetic features information. If there is no relative HR language, we look at the languages from the same family. Next, if there are several HR languages, we choose the most voluminous one sorted by the number of tokens. If no languages were found, the tuning is performed only for the second hypothesis.

- **Typological proximity**. To reveal donors in this case, we firstly cluster all collected languages by the WALS features. Secondly, we determine the donor as the closest HR language for the considered LR language. And the closeness is defined by the l_2 distance between LR and HR languages into the WALS features space.

According to the first approach, there were no HR languages found for such LR ones as: *'Atikamekw', 'Cheyenne', 'Greenlandic (South)', 'Quiché', 'Cherokee', 'Guaraní', 'Bambara', 'Yukaghir (Kolyma)'*. This is due to several reasons: 1) there are no genetically close languages collected, or they don't even exist; 2) genetically close languages are collected, but none of them are the HR ones.

4.2 Language Clustering

As for the second approach, the list of features was manually corrected. Some of the features were excluded as irrelevant to the transfer learning task: e.g., selective lexical features (*'138A Tea'* - whether the language has "tea" or "chai," only applied to some languages; *'129A Hand and Arm'*, *'130A Finger and Hand'*), an also region-specific features (*'10B Nasal Vowels in West Africa'*). The resulting list of the 192 selected features is presented in Appendix 3, consisting of grammatical features (shallow syntax can mostly influence lower levels of the model) and phonological features (can indirectly influence writing and tokenization).

Afterward, we applied K-Means as the clustering algorithm. The optimal number of clusters was chosen by the Silhouette method. We finally divided all 288 languages into 19 clusters. According to the conducted clustering, each cluster contained at least two languages (both low and HR types). Thereby if no candidates were found from the genetic proximity approach, we still have a donor from the typological proximity method. Finally, the donor language is determined as the closest one into the features space by l_2 distance.

[6] https://wals.info/feature

4.3 Transfer-learning experiments

In previous sections, we showed two different scenarios that we follow during the experiments. Also, we demonstrated two hypotheses to choose the HR donor languages for the experiments for each LR one. In this section, we describe the experiments in more detail.

In Appendix 2, you can see the table of Language statistics - a description of the collected dataset by languages in the numbers. In the transfer learning part, we use the data from HR donors and additionally train the model in the MLM setup. We limit the number of data of HR languages to train all the languages in the same conditions and training time: we chose 500k randomized sentence samples as the upper bound of the language data.

During the model training, we use the data of HR language in training and validation steps. The data of LR language is used only in the testing step to get the final score of how well the model understands the LR language.

4.3.1 Masked-Language modelling task and Perplexity

For model training, we chose the MLM task due to the structure of the collected data - we have the raw unlabeled textual data. According to the original procedure (Raffel et al., 2020) of training the mT5 model, we train the model to understand a new HR language on the denoising task by predicting masked spans (sequential set) of tokens.

In the MLM task, the Perplexity metric is used. We, as usual, define the Perplexity as the exponent of the cross-entropy loss:

$$\text{Perplexity}(X) = \exp\left\{-\frac{1}{t}\sum_{i}^{t}\log p_\theta(x_i|x_{<i})\right\}$$

where $\log p_\theta(x_i|x_{<i})$ is the log-likelihood of the i_{th} token conditioned on the preceding tokens $x_{<i}$ according to the model. As the final value of Perplexity, we use an average value by batches.

4.3.2 Types of measurements choice

During the experiments, we mainly get the Perplexity score of the model on LR languages before and after additional training on the related HR donor ones. And if we have no problem doing it at the first step, we can meet some problems at the second step. The problems occur due to the limitation of the data for LR languages.

The initial design of the experiments implied that we obtain the model checkpoints after each epoch of training and validation steps onto the HR donor language data. To prevent overfitting after a few epochs, we also proposed the second approach of getting checkpoints - save them after each 10% of the first epoch during the additional training. Looking ahead, we also tried early stopping, but it didn't lead to any improvements despite the second approach for measurements. Possible cause for that could be the limited amount of donors' data.

In the experiments, we used 14 GPU Tesla V100. Each epoch took $\sim$ 30 minutes by each GPU, and for each high-resource language, we trained the model in 20 epochs.

5 Results and Discussion

5.1 General model training results

As we discussed in 4.3.2, we have two approaches for getting the model's checkpoints. We wanted to figure out how often we need to save the model's state for obtaining the perplexity results on the low-resource languages. In Figure 2, you can see the examples of two different scopes of training. At the left figure, you can see that Perplexity is lower than the model's result in the zero-shot MLM task throughout 40% of the first epoch. At the right figure, perplexity values increase rapidly with the first epoch and are much higher than the zero-shot value. We also estimated how many languages achieved lower Perplexity after additional model training depending on the duration of this procedure. Results are as follows:

1. If we measure the results by the percentages of the first epoch, 95.7% experiments improved the results (lower Perplexity).

2. If we measure the results by the **epochs**, 42.7% experiments led to the perplexity decrement.

Here and below, we decided to show the results only with measurements during the first epoch since they are more successful according to the obtained numbers.

5.2 Results by languages

The results of mT5 additional training for each of the 46 LR languages are presented in Table 1. Each LR language is presented in the first column, while the donor languages are prescribed in the 3rd

LR language	LR perplexity before	HR language (genetic)	LR perplexity after training (minimum)	HR language (cluster)	LR perplexity after training (minimum)
Atikamekw	61.72	-	-	Asturian	**25.48**
Cheyenne	28.52	-	-	Asturian	**3.80**
Tsonga	40.41	Swahili	24.33	Asturian	**23.95**
Rundi	21.92	Swahili	**9.50**	Waray-Waray	10.32
Swati	40.65	Swahili	21.70	Belorussian	**16.66**
Kongo	26.46	Swahili	**4.66**	Aymara (Central)	8.35
Sesotho	12.77	Swahili	6.99	Kabiyé	**6.63**
Venda	31.95	Swahili	**11.89**	Kabiyé	13.82
Chichewa	13.72	Swahili	**8.80**	Malgwa	25.79
Kikuyu	38.81	Swahili	**8.41**	Xhosa	11.35
Chamorro	32.87	Cebuano	8.28	Tagalog	**7.41**
Cantonese	58.27	Mandarin	22.96	Bulgarian	**21.81**
Tok Pisin	30.81	Papiamentu	19.77	Afrikaans	**18.09**
Bislama	32.15	Papiamentu	17.54	Frisian (North)	**7.75**
Sranan	35.44	Papiamentu	**32.92**	Papiamentu	**32.92**
Coptic	4.72	Hebrew (Modern)	4.01	Sundanese	**3.69**
Greenlandic (South)	35.55	-	-	Mingrelian	**14.47**
Dagbani	47.81	Kabiyé	30.33	Sotho (Northern)	**30.06**
Bhojpuri	31.27	Hindi	19.11	Tulu	**14.08**
Romani (Lovari)	25.10	Hindi	**12.27**	Asturian	12.30
Kashmiri	26.27	Hindi	14.19	Malgwa	**13.60**
Kurmanji	32.44	Persian	**16.25**	Afrikaans	36.41
Tat (Muslim)	70.32	Persian	**45.45**	Jamaican Creole	66.91
Zhuang (Northern)	26.43	Thai	7.68	Bikol	**7.24**
Ewe	28.88	Swahili	**11.05**	Latvian	13.59
Akan	33.07	Swahili	**14.57**	Lao	16.28
Udi	55.01	Lezgian	96.51	Tulu	**14.00**
Tabassaran	57.19	Lezgian	73.16	Jamaican Creole	**51.75**
Tsakhur	41.74	Lezgian	36.94	Ladino	**14.04**
Madurese	33.61	Indonesian (Jakarta)	**14.34**	Frisian (Western)	19.08
Quiché	165.78	-	-	Asturian	**44.07**
Koryak	88.66	Chukchi	35.68	Tulu	**34.28**
Nadroga	34.94	Maori	**18.35**	Seediq	30.56
Nauruan	42.17	Maori	11.99	Shan	**3.70**
Samoan	12.52	Maori	**7.43**	Tongan	9.48
Tahitian	22.08	Maori	**6.29**	Tongan	9.82
Komi-Zyrian	110.02	Yazva	**32.98**	Kabiyé	70.13
Tigrinya	13.88	Hebrew (Modern)	7.50	Tulu	**7.36**
Cherokee	7.67	-	-	Minangkabau	**3.75**
Nanai	72.91	Solon	59.91	Papiamentu	**40.09**
GuaranÃ	3.99	-	-	Hausa	**2.90**
Shor	167.74	Chuvash	76.90	Altai (Southern)	**46.34**
Tofa	62.38	Chuvash	62.73	Jamaican Creole	**60.97**
Sango	23.20	Swahili	**8.19**	Hausa	13.60
Bambara	51.67	-	-	Navajo	**31.11**
Yukaghir (Kolyma)	**104.80**	-	-	Kannada	6785.49

Table 1: The original perplexity of mT5 model on the LR languages and their lowest perplexity after additional training on a HR donor language. In the "HR language (relative)" column, the omissions are explained by the fact that among the 288 languages we collected, there was no HR language found from the same genus or family as the LR language. See the Appendix A.6 for the training perplexity curves for the described languages and candidates.

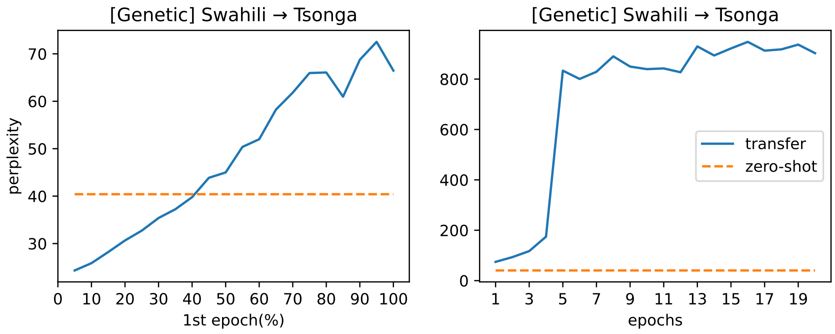


Figure 2: Example of additional training mT5-Base on Swahili and measuring perplexity by epochs and within the first epoch.

and 5th columns. The minimal perplexity obtained while additional training is presented in columns 2 and 4. Basing on the results, the best language pairs for transfer learning are:

- by the genetic proximity: *Swahili → Rundi, Swahili → Kongo, Swahili → Venda, Swahili → Chichewa, Swahili → Kikuyu, Swahili → Ewe, Swahili → Akan, Swahili → Sango, Papiamentu → Sranan, Hindi → Romani (Lovari), Persian → Kurmanji, Persian → Tat (Muslim), Indonesian (Jakarta) → Madurese, Maori → Nadroga, Maori → Samoan, Maori → Tahitian, Yazva → Komi-Zyrian.*

- by the typological proximity: *Asturian → Atikamekw, Asturian → Cheyenne, Asturian → Tsonga, Asturian → Quiché, Belorussian → Swati, Kabiyé → Sesotho, Tagalog → Chamorro, Bulgarian → Cantonese, Afrikaans → Tok Pisin, Frisian (North) → Bislama, Papiamentu → Sranan, Papiamentu → Nana, Sundanese → Coptic, Mingrelian → Greenlandic (South), Sotho (Northern) → Dagbani, Tulu → Bhojpuri, Tulu → Koryak, Tulu → Udi, Tulu → Tigrinya, Malgwa → Kashmiri, Bikol → Zhuang (Northern), Jamaican Creole → Tabassaran, Jamaican Creole → Tofa, Ladino → Tsakhur, Shan → Nauruan, Minangkabau → Cherokee, Hausa → GuaranÃ, Altai (Southern) → Shor, Navajo → Bambara.*

Modelling Yukaghir (Kolyma) language with different donors does not show any improvements. Swahili appears to be the most popular and generally good donor for its genus, and tools and models for the Bantoid languages can be easily developed with transfer learning. The same applies to the Maori language and the Oceanic genus.

We recognize that typological (clustering) language pairs are not obvious. The most important features of the clustering (Appendix 4) seem to be universally important in terms of syntax. As numerous works on probing (Ravishankar et al., 2019; Mikhailov et al., 2021) show, syntactic information is stored in the embeddings of intermediate layers of transformer models and can affect the additional training process.

According to the Endangered Languages project [7], such languages from the Table 1 are

1. *critically endangered*: Tat (Muslim) (perplexity lowered from 70.32 to 45.45), Nanai (72.91 → 40.09), Shor (167.74 → 46.34), Tofa (62.3 → 60.97), Yukaghir (Kolyma) (104.8 → not lowered);

2. *endangered*: Cheyenne (28.52 → 3.80), Udi (55.01 → 14.00), Koryak (88.66 → 34.28);

3. *threatened*: Nauruan (42.17 → 3.70), Cherokee (7.67 → 3.75);

4. *vulnerable or at risk*: Atikamekw (61.72 → 25.48), Chamorro (32.87 → 7.41), Romani (Lovari)(25.10 → 12.27), Tabassaran (57.19 → 51.75), Greenlandic (South) (35.55 → 14.47), Tsakhur (41.74 → 14.04), Quiché (165.78 → 44.07).

For many of the above-mentioned languages, a significant decrease in perplexity is noticeable.

5.3 General Observations and Discussion

In future research, we plan to further elaborate on the results of the provided method. Among other things, we are planning to explore the perplexity

[7] https://endangeredlanguages.com/

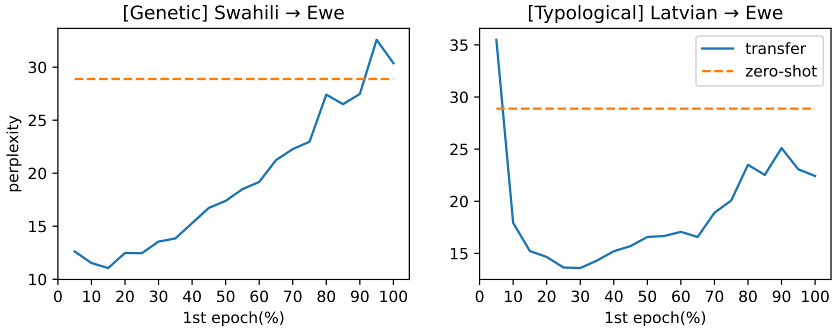


Figure 3: Example for the case of Ewe language (LR): additional training on the relative language (Swahili) and feature-related language from k-MEANS (Latvian)

scores with more classical language pairs and similar experimental setups, and also conduct probing experiments on the model checkpoints.

The data pipeline we provided can easily reproduce the multilingual corpus collection; it can also be extended by external resources. We would like to include the JW300 (Agić and Vulić, 2019) case as well, but its online resources were not available at the time of research. As a continuation of the approach, corresponding experiments with downstream tasks can be conducted for cross-lingual transfer evaluation. Unfortunately, data for such tasks is only available for a very limited number of languages[8].

An interesting observation is the fact that the overall percentage of the successful additional training experiments, lowering perplexity for the target LR language, is mostly based on the clustering hypothesis, not the genus relation.

The clustering-based choice of the HR donor language also reveals a question for several separate languages: the following pairs *Ladino → Tsakhur, Latvian → Ewe, Lao → Akan, Xhosa → Kikuyu, Tulu → Koryak, Tongan → Tahitian, Altai (Southern) → Shor* show an atypical picture in which tuning a model in an unrelated language with each iteration systematically reduces perplexity in a LR language.

It remains a question for future research, why exactly these pairs of languages are so successful and whether these results are arguments in favour of typological similarity being more important than genetic relationship for the transfer learning problem. However, this approach shows its practical applicability, and for many languages that do not have closely related multi-resource languages, it can become the principal way to transfer knowledge.

6 Conclusion

We present several methods for transfer learning in language modelling based on training a donor high-resource language pair for a low-resource language. The methods work both for related language pairs based on the same language genus or family and unrelated pairs with similar grammatical features forming the same clusters.

The results show that the proposed transfer learning techniques achieve lower perplexity on the target low-resource languages even if no data is available for training but only for evaluation on the test set. The result of our experiments shows that transfer learning is possible not only between related languages but also often turns out to be possible on the basis of the typological similarity of languages, with syntactic features being the key ones.

In that way, we point at the list of 46 new languages not included in the NLP practice but having a non-zero text representation online. We present new pairs of languages between which the transfer of knowledge is possible.

We believe that the overall multilingual prospects of language technology can be much more bright if more effort is put into the development and standardization of available field linguistics resources and their incorporation into language modelling development. [9]

[8] for example, Universal Dependencies data, XGLUE, and XTREME benhcmarks

[9] The multilingual data is available at shorturl.at/ikmI9 (anonymous). The code of the experiments is attached to this submission as a zip-archive.

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Anna Belew. 2021. The endangered languages project (elp): Collaborative infrastructure and knowledge-sharing to support indigenous and endangered languages.

Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. 2016. Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549.

Marie-Catherine de Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2021. Ethnologue: Languages of the world. twenty-fourth edition.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 138–147.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2020. A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309*.

Katharina Kann, Ophélie Lacroix, and Anders Søgaard. 2020. Weakly supervised pos taggers perform poorly on truly low-resource languages. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8066–8073.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

Irina Krylova, Boris Orekhov, Ekaterina Stepanova, and Lyudmila Zaydelman. 2015. Languages of russia: Using social networks to collect texts. In *Russian summer school in information retrieval*, pages 179–185. Springer.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020a. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv*, abs/2004.01401.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020b. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv preprint arXiv:2004.01401*.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert?

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Vladislav Mikhailov, Oleg Serikov, and Ekaterina Artemova. 2021. Morph call: Probing morphosyntactic content of multilingual transformers. *arXiv preprint arXiv:2104.12847*.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Vinit Ravishankar, Memduh Gökırmak, Lilja Øvrelid, and Erik Velldal. 2019. Multilingual probing of deep pre-trained contextual encoders. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 37–47.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Graham Neubig, and Melvin Johnson. 2021. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. *arXiv preprint arXiv:2104.07412*.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert?

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Ze Yang, Wei Wu, Jian Yang, Can Xu, and Zhoujun Li. 2019. Low-resource response generation with template prior. 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 1886–1897, Hong Kong, China. Association for Computational Linguistics.

A Appendix

A.1 Appendix 1. All the languages present in the corpus

288 languages with available textual data, which are: *'Abaza', 'Acehnese', 'Arabic (Egyptian)', 'Afrikaans', 'Akan', 'Albanian', 'Amharic', 'Arabic (Moroccan)', 'Arabic (Modern Standard)', 'Apurinã', 'Archi', 'Arabic (Lebanese)', 'Armenian (Eastern)', 'Armenian (Western)', 'Armenian (Iranian)', 'Adyghe (Shapsugh)', 'Altai (Southern)', 'Assamese', 'Asturian', 'Atayal', 'Atikamekw', 'Avar', 'Awadhi', 'Aymara (Central)', 'Azerbaijani', 'Azari (Iranian)', 'Balinese', 'Bambara', 'Beja', 'Bengali', 'Bhojpuri', 'Bikol', 'Belorussian', 'Breton', 'Burmese', 'Bashkir', 'Bislama', 'Basque', 'Bugis', 'Bulgarian', 'Buriat', 'Choctaw', 'Cebuano', 'Chamorro', 'Chechen', 'Cherokee', 'Chukchi', 'Chuvash', 'Chichewa', 'Cantonese', 'Coptic', 'Crimean Tatar', 'Cornish', 'Catalan', 'Chatino (Yaitepec)', 'Cheyenne', 'Czech', 'Dagbani', 'Dogri', 'Dhivehi', 'Dargwa', 'Danish', 'Dutch', 'Dutch (Zeeuws)', 'English', 'Estonian', 'Even', 'Ewe', 'Faroese', 'Finnish', 'Frisian (North)', 'French', 'Frisian', 'Frisian (Western)', 'Fuzhou', 'Gaelic (Scots)', 'Gagauz', 'Georgian', 'German', 'Guianese French Creole', 'Gilaki', 'Guajajara', 'Galician', 'Greek (Modern)', 'German (Ripuarian)', 'Gorontalo', 'Greenlandic (South)', 'German (Timisoara)', 'Guaraní', 'Gujarati', 'German (Viennese)', 'German (Zurich)', 'Hakka', 'Hausa', 'Hawaiian', 'Haitian Creole', 'Hebrew (Modern)', 'Hindi', 'Hungarian', 'Icelandic', 'Igbo', 'Ilocano', 'Indonesian', 'Ingush', 'Indonesian (Jakarta)', 'Irish', 'Irish (Munster)', 'Italian', 'Itelmen', 'Italian (Genoa)', 'Italian (Napolitanian)', 'Italian (Turinese)', 'Javanese', 'Jamaican Creole', 'Japanese', 'Kabardian', 'Kashmiri', 'Kazakh', 'Kabyle', 'Kabiyé', 'Kirghiz', 'Khakas', 'Khmer', 'Kikuyu', 'Kinyarwanda', 'Kurmanji', 'Karakalpak', 'Kannada', 'Kongo', 'Komi-Permyak', 'Korean', 'Kapampangan', 'Karachay-Balkar', 'Kurdish (Central)', 'Karelian', 'Koryak', 'Khanty', 'Kumyk', 'Komi-Zyrian', 'Lak', 'Lao', 'Latvian', 'Luganda', 'Ladin', 'Lezgian', 'Low German', 'Lingala', 'Lithuanian', 'Liv', 'Ladino', 'Luxemburgeois', 'Mari (Hill)', 'Maithili', 'Maori', 'Macedonian', 'Madurese', 'Meithei', 'Mingrelian', 'Marathi', 'Minangkabau', 'Mongol (Khamnigan)', 'Malgwa', 'Maltese', 'Malay', 'Mari (Meadow)', 'Mordvin (Moksha)', 'Mandarin', 'Mansi', 'Manx', 'Mordvin (Erzya)', 'Mon', 'Marshallese', 'Mundurukú', 'Malayalam', 'Mazanderani', 'Nanai', 'Nauruan', 'Navajo', 'Ndonga', 'Nadroga', 'Nepali', 'Nias', 'Norwegian', 'Narom', 'Neo-Aramaic (Assyrian)', 'Nenets (Tundra)', 'Nivkh (South Sakhalin)', 'Newar (Dolakha)', 'Oirat', 'Ossetic', 'Oriya', 'Panjabi', 'Papiamentu', 'Pangasinan', 'Polish', 'Portuguese', 'Provençal', 'Persian', 'Qafar', 'Quiché', 'Romani (Lovari)', 'Rundi', 'Romanian', 'Romansch (Sursilvan)', 'Russian', 'Rutul', 'Samoan', 'Sango', 'Serbian-Croatian', 'Sindhi', 'Seediq', 'Sesotho', 'Shan', 'Shona', 'Shor', 'Slovene', 'Seminole', 'Sinhala', 'Saami (Northern)', 'Solon', 'Somali', 'Sorbian (Upper)', 'Spanish', 'Sranan', 'Sardinian', 'Sorbian (Lower)', 'Santali', 'Sotho (Northern)', 'Sundanese', 'Slovincian', 'Slovak', 'Swahili', 'Swedish', 'Swati', 'Swedish (Västerbotten)', 'Tagalog', 'Tahitian', 'Tajik', 'Tashlhiyt', 'Tabassaran', 'Telugu', 'Thai', 'Tigrinya', 'Turkmen', 'Tamil', 'Tibetan (Modern Literary)', 'Tat (Muslim)', 'Tongan', 'Tofa', 'Tok Pisin', 'Tsakhur',*

'Tsonga', 'Tamil (Spoken)', 'Tswana', 'Tetun', 'Tulu', 'Tupi', 'Turkish', 'Tuvan', 'Tatar', 'Udi', 'Udmurt', 'Ukrainian', 'Urdu', 'Urubú-Kaapor', 'Uyghur', 'Uzbek', 'Venda', 'Veps', 'Vietnamese', 'Welsh', 'Wolof', 'Warlpiri', 'Wu', 'Waray-Waray', 'Xhosa', 'Yi', 'Yukaghir (Kolyma)', 'Yakut', 'Yiddish (Lithuanian)', 'Yoruba', "Yup'ik (Central)", 'Yurt Tatar', 'Yukaghir (Tundra)', 'Yazva', 'Zazaki', 'Zhuang (Northern)', 'Zulu'.

A.2 Appendix 2. Language statistics

Number of tokens for each language is counted with the usage of the mT5-Base tokenizer.

Name	N_tokens, kk	N_symbols, kk	Name	N_tokens, kk	N_symbols, kk
Abaza	1.33	2.35	Buriat	172.16	344.66
Acehnese	1.47	3.09	Choctaw	0.001	0.002
Arabic (Egyptian)	157.47	291.76	Cebuano	1365.37	3319.73
Afrikaans	46.5	126.33	Chamorro	0.06	0.13
Akan	0.33	0.62	Chechen	378.15	477.3
Albanian	0.002	0.005	Cherokee	0.17	0.22
Amharic	5.37	6.85	Chukchi	4.08	5.12
Arabic (Moroccan)	1.1	2.07	Chuvash	277.48	442.77
Arabic (Modern Standard)	1.83	1.83	Chichewa	0.28	0.75
Apurinã	0.002	0.0035	Cantonese	0.02	0.02
Archi	0.0012	0.0019	Coptic	0.1	0.13
Arabic (Lebanese)	0.0015	0.0026	Crimean Tatar	1.24	2.6
Armenian (Eastern)	0.12	0.26	Cornish	0.9	1.88
Armenian (Western)	0.09	0.17	Catalan	463.18	1168.29
Armenian (Iranian)	212.94	569.39	Chatino (Yaitepec)	1.21	3.05
Adyghe (Shapsugh)	3.32	5.58	Cheyenne	0.06	0.1
Altai (Southern)	2.46	4.6	Czech	336.81	819.13
Assamese	11.18	18.87	Dagbani	0.28	0.55
Asturian	117.6	304.59	Dogri	0.006	0.009
Atayal	0.71	1.35	Dhivehi	4.35	5.77
Atikamekw	0.33	0.71	Dargwa	11.64	23.4
Avar	5.73	10.82	Danish	0.52	1.46
Awadhi	0.57	1.04	Dutch	598	1675.76
Aymara (Central)	0.92	1.81	Dutch (Zeeuws)	1.43	3.12
Azerbaijani	90.5	230.74	English	7920.93	24002.62
Azari (Iranian)	39.73	74.9	Estonian	97.8	265.66
Balinese	2.04	4.87	Even	0	0.01
Bambara	0.17	0.31	Ewe	0.085	0.153
Beja	0.003	0.003	Faroese	4.17	9.4
Bengali	28.14	56.44	Finnish	243.97	715.36
Bhojpuri	0.01	0.02	Frisian (North)	2.74	5.65
Bikol	3.39	8.63	French	1541.64	4046.63
Belorussian	168.71	467.04	Frisian	0.007	0.016
Breton	21.79	43.06	Frisian (Western)	29.12	69.47
Burmese	26.12	64.79	Fuzhou	1.95	2.86
Bashkir	58.9	122.15	Gaelic (Scots)	4.52	9.25
Bislama	0.13	0.28	Gagauz	0.43	1.02
Basque	12.73	35.51	Georgian	65.7	149.49
Bugis	0.98	2.05	German	6.94	21.3
Bulgarian	147.44	367.2	Guianese French Creole	0.53	1.06

Name	N_tokens, kk	N_symbols, kk	Name	N_tokens, kk	N_symbols, kk	Name	N_tokens, kk	N_symbols, kk
Gilaki	1.33	2.38	Kinyarwanda	0.71	1.65	Samoan	0.31	0.61
Guajajara	0.0016	0.0023	Kurmanji	0.02	0.04	Sango	0.04	0.06
Galician	108.96	282.46	Karakalpak	0.71	1.55	Serbian-Croatian	375.92	828.15
Greek (Modern)	167.59	387.4	Kannada	40.49	94.87	Sindhi	9.28	15
German (Ripuarian)	1.01	2.21	Kongo	0.12	0.26	Seediq	1.14	2.28
Gorontalo	1.3	3.1	Komi-Permyak	1.82	3.19	Sesotho	0.22	0.49
Greenlandic (South)	0.15	0.36	Korean	254.45	318.2	Shan	5.09	7.59
German (Timisoara)	1761.41	5469.18	Kapampangan	1.85	4.41	Shona	1.32	3.11
Guaraní	0.03	0.02	Karachay-Balkar	4.38	9.14	Shor	0.18	0.31
Gujarati	19.35	34.67	Kurdish (Central)	16.77	29.46	Slovene	92.81	239
German (Viennese)	9.44	21.27	Karelian	0.01	0.02	Seminole	0	0
German (Zurich)	0.003	0.006	Koryak	0.25	0.43	Sinhala	18.78	37.48
Hakka	1.67	2.68	Khanty	0.0004	0.0005	Saami (Northern)	1.27	2.62
Hausa	5.58	13.62	Kumyk	1.16	2.45	Solon	1.49	2.93
Hawaiian	0.53	1.02	Komi-Zyrian	0.03	0.05	Somali	3.51	8.41
Haitian Creole	7.72	15.97	Lak	16.46	30.72	Sorbian (Upper)	3.71	7.64
Hebrew (Modern)	308.46	623.18	Lao	1.74	4.17	Spanish	1233.15	3408.23
Hindi	81.33	160.75	Latvian	54.26	135.41	Sranan	0.2	0.43
Hungarian	297.86	779.47	Luganda	1.47	3.38	Sardinian	3.39	7.95
Icelandic	23.58	55.33	Ladin	0.45	0.94	Sorbian (Lower)	0.83	1.69
Igbo	1.16	2.27	Lezgian	10.34	18.87	Santali	6.21	8.12
Ilocano	4.39	10.03	Low German	20.84	51.3	Sotho (Northern)	0.84	1.86
Indonesian	0.28	0.89	Lingala	0.53	1.06	Sundanese	12.54	31.15
Ingush	8.09	14.27	Lithuanian	78.83	199.57	Slovincian	1.2	2.14
Indonesian (Jakarta)	209.58	612.28	Liv	0.004	0.009	Slovak	91.75	224.02
Irish	0.27	0.6	Ladino	1.31	3.25	Swahili	14.32	35.39
Irish (Munster)	15.82	34.2	Luxemburgeois	17.24	41.58	Swedish	0.36	0.99
Italian	1018.44	2776.36	Mari (Hill)	1.71	2.91	Swati	0.14	0.34
Itelmen	0.0005	0.0008	Maithili	2.86	5.27	Swedish (Västerbotten)	882.57	2204.72
Italian (Genoa)	2.57	4.9	Maori	1.2	2.25	Tagalog	21.96	54.45
Italian (Napolitanian)	1.99	3.9	Macedonian	83.34	211.18	Tahitian	0.11	0.19
Italian (Turinese)	12.27	23.48	Madurese	0.2	0.42	Tajik	20.91	43.76
Javanese	17.44	43.98	Meithei	0.47	0.94	Tashlhiyt	0.53	0.89
Jamaican Creole	0.42	0.9	Mingrelian	4.43	8.19	Tabassaran	0.06	0.11
Japanese	750.08	1244.1	Marathi	17.82	36.2	Telugu	79.39	178.59
Kabardian	18.8	29.62	Minangkabau	34.7	86.25	Thai	79.96	226.41
Kashmiri	0.13	0.22	Mongol (Khamnigan)	13.48	30.69	Tigrinya	0.13	0.14
Kazakh	72.64	184.75	Malgwa	21.64	48.87	Turkmen	4.05	8.47
Kabyle	1.58	2.83	Maltese	5.37	11.79	Tamil	0.03	0.08
Kabiyé	1.84	2.39	Malay	83.7	241.85	Tibetan (Modern Literary)	34.98	41.26
Galician	108.96	282.46	Mari (Meadow)	189.06	275.65	Tat (Muslim)	0.06	0.11
Greek (Modern)	167.59	387.4	Mordvin (Moksha)	1.45	2.24	Tongan	0.36	0.65
German (Ripuarian)	1.01	2.21	Mandarin	497.71	659.72	Tofa	0.03	0.06
Gorontalo	1.3	3.1	Mansi	2.58	3.99	Tok Pisin	0.16	0.34
Greenlandic (South)	0.15	0.36	Manx	1.57	3.07	Tsakhur	0.09	0.15
German (Timisoara)	1761.41	5469.18	Mordvin (Erzya)	7.81	15.25	Tsonga	0.25	0.58
Guaraní	0.03	0.02	Mon	4.98	7.33	Tamil (Spoken)	65.88	188.99
Gujarati	19.35	34.67	Marshallese	0.002	0.003	Tswana	0.5	1.15
German (Viennese)	9.44	21.27	Mundurukú	0.002	0.002	Tetun	0.42	1.01
German (Zurich)	0.003	0.006	Malayalam	45.49	118.33	Tulu	1.14	2.15
Hakka	1.67	2.68	Mazanderani	2.73	5.08	Tupi	0.006	0.008
Hausa	5.58	13.62	Nanai	0.24	0.41	Turkish	169.37	468.8
Hawaiian	0.53	1.02	Nauruan	0.13	0.26	Tuvan	24.8	51.9
Haitian Creole	7.72	15.97	Navajo	5.67	8.52	Tatar	59.84	127.89
Hebrew (Modern)	308.46	623.18	Ndonga	0.003	0.008	Udi	0.1	0.16
Hindi	81.33	160.75	Nadroga	0.2	0.43	Udmurt	4.76	9.24
Hungarian	297.86	779.47	Nepali	13.5	27.54	Ukrainian	619.13	1523.74
Icelandic	23.58	55.33	Nias	0.36	0.77	Urdu	57.23	110.15
Igbo	1.16	2.27	Norwegian	224.43	608.76	Urubú-Kaapor	0.001	0.001
Ilocano	4.39	10.03	Narom	0.98	1.96	Uyghur	12.66	18.24
Indonesian	0.28	0.89	Neo-Aramaic (Assyrian)	0.0026	0.0021	Uzbek	45.15	104.75
Ingush	8.09	14.27	Nenets (Tundra)	0.47	0.78	Venda	0.11	0.25
Indonesian (Jakarta)	209.58	612.28	Nivkh (South Sakhalin)	0.73	1.14	Veps	2.96	6.83
Irish	0.27	0.6	Newar (Dolakha)	24.93	45.29	Vietnamese	424.35	742.98
Irish (Munster)	15.82	34.2	Oirat	7.97	14.12	Welsh	40.94	82.34
Italian	1018.44	2776.36	Ossetic	2.96	4.91	Wolof	1.3	2.54
Itelmen	0.0005	0.0008	Oriya	16.53	18.51	Warlpiri	0	0
Italian (Genoa)	2.57	4.9	Panjabi	27.68	42.51	Wu	6.67	9.11
Italian (Napolitanian)	1.99	3.9	Papiamentu	11.04	19.5	Waray-Waray	187.45	446.09
Italian (Turinese)	12.27	23.48	Pangasinan	0.63	1.29	Xhosa	0.61	1.56
Javanese	17.44	43.98	Polish	629.24	1616.18	Yi	0.001	0.001
Jamaican Creole	0.42	0.9	Portuguese	600.72	1550.9	Yukaghir (Kolyma)	0.026	0.044
Japanese	750.08	1244.1	Provençal	32.42	76.25	Yakut	16.84	33.18
Kabardian	18.8	29.62	Persian	298.12	601.61	Yiddish (Lithuanian)	7.43	14.58
Kashmiri	0.13	0.22	Qafar	0	0.0001	Yoruba	4.69	8.03
Kazakh	72.64	184.75	Quiché	0.02	0.04	Yup'ik (Central)	0.004	0.009
Kabyle	1.58	2.83	Romani (Lovari)	0.2	0.49	Yurt Tatar	2.28	4.04
Kabiyé	1.84	2.39	Rundi	0.14	0.31	Yukaghir (Tundra)	0	0.01
Kirghiz	27.32	65.52	Romanian	195.88	485.41	Yazva	14.23	25.9
Khakas	1.44	2.89	Romansch (Sursilvan)	5.07	12.59	Zazaki	5.46	10.8
Khmer	10.98	28.84	Russian	2372.32	6397.2	Zhuang (Northern)	0.28	0.52
Kikuyu	0.18	0.34	Rutul	0.36	0.67	Zulu	1	2.34

A.3 Appendix 3. The features used in the language clusterization

The resulting list of features used in the clusterization: *'1A Consonant Inventories', '2A Vowel Quality Inventories', '3A Consonant-Vowel Ratio', '4A Voicing in Plosives and Fricatives', '5A Voicing and Gaps in Plosive Systems', '6A Uvular Consonants', '7A Glottalized Consonants', '8A Lateral Consonants', '9A The Velar Nasal', '10A Vowel Nasalization', '11A Front Rounded Vowels', '12A Syllable Structure', '13A Tone', '14A Fixed Stress Locations', '15A Weight-Sensitive Stress', '16A Weight Factors in Weight-Sensitive Stress Systems', '17A Rhythm Types', '18A Absence of Common Consonants', '19A Presence of Uncommon Consonants', '20A Fusion of Selected Inflectional Formatives', '21A Exponence of Selected Inflectional Formatives', '22A Inflectional Synthesis of the Verb', '23A Locus of Marking in the Clause', '24A Locus of Marking in Possessive Noun Phrases', '25A Locus of Marking: Whole-language Typology', '26A Prefixing vs. Suffixing in Inflectional Morphology', '27A Reduplication', '28A Case Syncretism', '29A Syncretism in Verbal Person/Number Marking', '30A Number of Genders', '31A Sex-based and Non-sex-based Gender Systems', '32A Systems of Gender Assignment', '33A Coding of Nominal Plurality', '34A Occurrence of Nominal Plurality', '35A Plurality in Independent Personal Pronouns', '36A The Associative Plural', '37A Definite Articles', '38A Indefinite Articles', '39A Inclusive/Exclusive Distinction in Independent Pronouns', '40A Inclusive/Exclusive Distinction in Verbal Inflection', '41A Distance Contrasts in Demonstratives', '42A Pronominal and Adnominal Demonstratives', '43A Third Person Pronouns and Demonstratives', '44A Gender Distinctions in Independent Personal Pronouns', '45A Politeness Distinctions in Pronouns', '46A Indefinite Pronouns', '47A Intensifiers and Reflexive Pronouns', '48A Person Marking on Adpositions', '49A Number of Cases', '50A Asymmetrical Case-Marking', '51A Position of Case Affixes', '52A Comitatives and Instrumentals', '53A Ordinal Numerals', '54A Distributive Numerals', '55A Numeral Classifiers', '56A Conjunctions and Universal Quantifiers', '57A Position of Pronominal Possessive Affixes', '58A Obligatory Possessive Inflection', '59A Possessive Classification', '60A Genitives, Adjectives and Relative Clauses', '61A Adjectives without Nouns', '62A Action Nominal Constructions', '63A Noun Phrase Conjunction', '64A Nominal and Verbal Conjunction', '65A Perfective/Imperfective Aspect', '66A The Past Tense', '67A The Future Tense', '68A The Perfect', '69A Position of Tense-Aspect Affixes', '70A The Morphological Imperative', '71A The Prohibitive', '72A Imperative-Hortative Systems', '73A The Optative', '74A Situational Possibility', '75A Epistemic Possibility', '76A Overlap between Situational and Epistemic Modal Marking', '77A Semantic Distinctions of Evidentiality', '78A Coding of Evidentiality', '79A Suppletion According to Tense and Aspect', '80A Verbal Number and Suppletion', '81A Order of Subject, Object and Verb', '82A Order of Subject and Verb', '83A Order of Object and Verb', '84A Order of Object, Oblique, and Verb', '85A Order of Adposition and Noun Phrase', '86A Order of Genitive and Noun', '87A Order of Adjective and Noun', '88A Order of Demonstrative and Noun', '89A Order of Numeral and Noun', '90A Order of Relative Clause and Noun', '91A Order of Degree Word and Adjective', '92A Position of Polar Question Particles', '93A Position of Interrogative Phrases in Content Questions', '94A Order of Adverbial Subordinator and Clause', '95A Relationship between the Order of Object and Verb and the Order of Adposition and Noun Phrase', '96A Relationship between the Order of Object and Verb and the Order of Relative Clause and Noun', '97A Relationship between the Order of Object and Verb and the Order of Adjective and Noun', '98A Alignment of Case Marking of Full Noun Phrases', '99A Alignment of Case Marking of Pronouns', '100A Alignment of Verbal Person Marking', '101A Expression of Pronominal Subjects', '102A Verbal Person Marking', '103A Third Person Zero of Verbal Person Marking', '104A Order of Person Markers on the Verb', "105A Ditransitive Constructions: The Verb 'Give'", '106A Reciprocal Constructions', '107A Passive Constructions', '108A Antipassive Constructions', '109A Applicative Constructions', '110A Periphrastic Causative Constructions', '111A Nonperiphrastic Causative Constructions', '112A Negative Morphemes', '113A Symmetric and Asymmetric Standard Negation', '114A Subtypes of Asymmetric Standard Negation', '115A Negative Indefinite Pronouns and Predicate Negation', '116A Polar Questions', '117A Predicative Possession', '118A Predicative Adjectives', '119A Nominal and Locational Predication', '120A Zero Copula for Predicate Nominals', '121A Comparative Constructions', '122A Relativization* *on Subjects', '123A Relativization on Obliques', "124A 'Want' Complement Subjects", '125A Purpose Clauses', "126A 'When' Clauses", '127A Reason Clauses', '128A Utterance Complement Clauses', '129A Hand and Arm', '130A Finger and Hand', '131A Numeral Bases', '132A Number of Non-Derived Basic Colour Categories', '133A Number of Basic Colour Categories', '134A Green and Blue', '135A Red and Yellow', '136A M-T Pronouns', '137A N-M Pronouns', '138A Tea', '139A Irregular Negatives in Sign Languages', '140A Question Particles in Sign Languages', '141A Writing Systems', '142A Para-Linguistic Usages of Clicks', '143F Postverbal Negative Morphemes', '90B Prenominal relative clauses', '144Y The Position of Negative Morphemes in Object-Initial Languages', '90C Postnominal relative clauses', '144P NegSOV Order', '144J SVNegO Order', '144N Obligatory Double Negation in SOV languages', '144S SOVNeg Order', '144X Verb-Initial with Clause-Final Negative', '144A Position of Negative Word With Respect to Subject, Object, and Verb', '90G Double-headed relative clauses', '90E Correlative relative clauses', '144V Verb-Initial with Preverbal Negative', '144I SNegVO Order', '144R SONegV Order', '143B Obligatory Double Negation', '144M Multiple Negative Constructions in SOV Languages', '144U Double negation in verb-initial languages', '144G Optional Double Negation in SVO languages', '144K SVONeg Order', '144B Position of negative words relative to beginning and end of clause and with respect to adjacency to verb', '144F Obligatory Double Negation in SVO languages', '90D Internally-headed relative clauses', '144E Multiple Negative Constructions in SVO Languages', '144D The Position of Negative Morphemes in SVO Languages', '81B Languages with two Dominant Orders of Subject, Object, and Verb', '143E Preverbal Negative Morphemes', '143C Optional Double Negation', '90F Adjoined relative clauses', '143A Order of Negative Morpheme and Verb', '144W Verb-Initial with Negative that is Immediately Postverbal or between Subject and Object', '144O Optional Double Negation in SOV languages', '144Q SNegOV Order', '144L The Position of Negative Morphemes in SOV Languages', '144H NegSVO Order', '144C Languages with different word order in negative clauses', '144T The Position of Negative Morphemes in Verb-Initial Languages', '143G Minor morphological means of signaling negation', '143D Optional Triple Negation', '39B Inclusive/Exclusive Forms in Pama-Nyungan', '137B M in Second Person Singular', '136B M in First Person Singular', '109B Other Roles of Applied Objects', '10B Nasal Vowels in West Africa', '25B Zero Marking of A and P Arguments', '21B Exponence of Tense-Aspect-Mood Inflection', '108B Productivity of the Antipassive Construction', "130B Cultural Categories of Languages with Identity of 'Finger' and 'Hand'", '58B Number of Possessive Nouns', '79B Suppletion in Imperatives and Hortatives'*

A.4 Appendix 4. Languages clustering vizualization

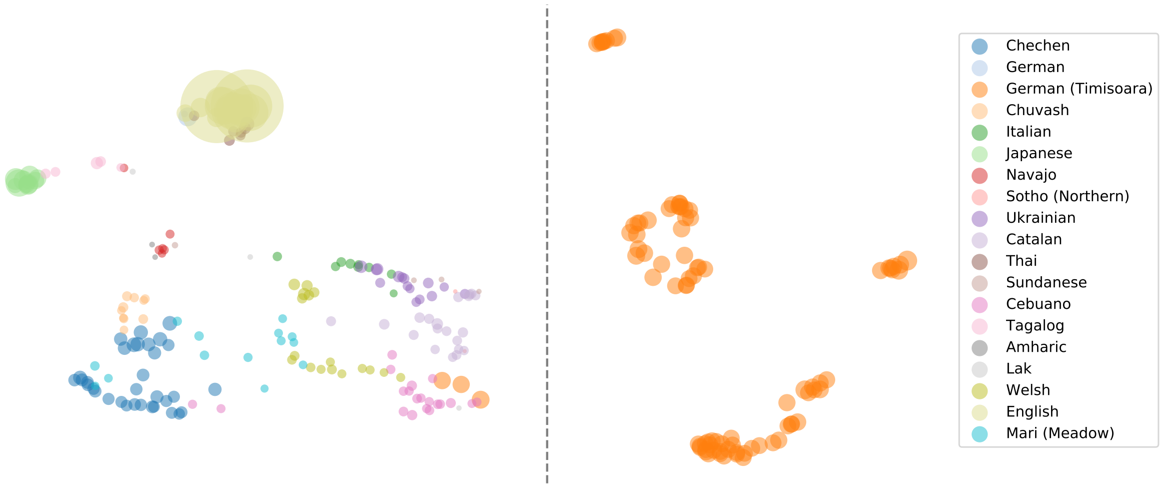


Figure 4: UMAP-vizualization of languages clustering based on the WALS features. As cluster labels, the most resourced languages of clusters are used. An orange cluster on the right is actually located much further to the right. Dots are languages, size encodes the (amount of data in corpus / number of UNK WALS features) ratio.

A.4.1 List of the "garbage" cluster languages

. The hugest cluster (see Figure 4) which tends to have modular structure itself, is composed of the following languages:

Abaza, Adyghe (Shapsugh), Afrikaans, Altai (Southern), Arabic (Lebanese), Armenian (Iranian), Asturian, Atikamekw, Awadhi, Balinese, Bislama, Cheyenne, Crimean Tatar, Dogri, Dutch (Zeeuws), Faroese, Frisian (North), Frisian (Western), Fuzhou, Gagauz, Galician, German (Ripuarian), German (Viennese), German (Zurich), Gilaki, Gorontalo, Greenlandic (South), Guianese French Creole, Haitian Creole, Indonesian (Jakarta), Irish (Munster), Italian (Genoa), Italian (Napolitanian), Italian (Turinese), Jamaican Creole, Javanese, Karelian, Kazakh, Khakas, Kumyk, Kurmanji, Ladin, Ladino, Liv, Low German, Luxemburgeois, Madurese, Malay, Maltese, Manx, Mari (Hill), Marshallese, Mazanderani, Mingrelian, Mordvin (Moksha), Nanai, Narom, Nenets (Tundra), Neo-Aramaic (Assyrian), Nivkh (South Sakhalin), Papiamentu, Provençal, Quiché, Romani (Lovari), Rundi, Shan, Shor, Sindhi, Slovak, Slovincian, Solon, Sorbian (Lower), Sorbian (Upper), Sranan, Swedish (Västerbotten), Tabassaran, Tamil (Spoken), Tat (Muslim), Tofa, Tok Pisin, Tsakhur, Tsonga, Tswana, Tupi, Veps, Waray-Waray, Wu, Yazva, Yiddish (Lithuanian), Yurt Tatar .

A.5 Appendix 5. Feature importance for language clustering

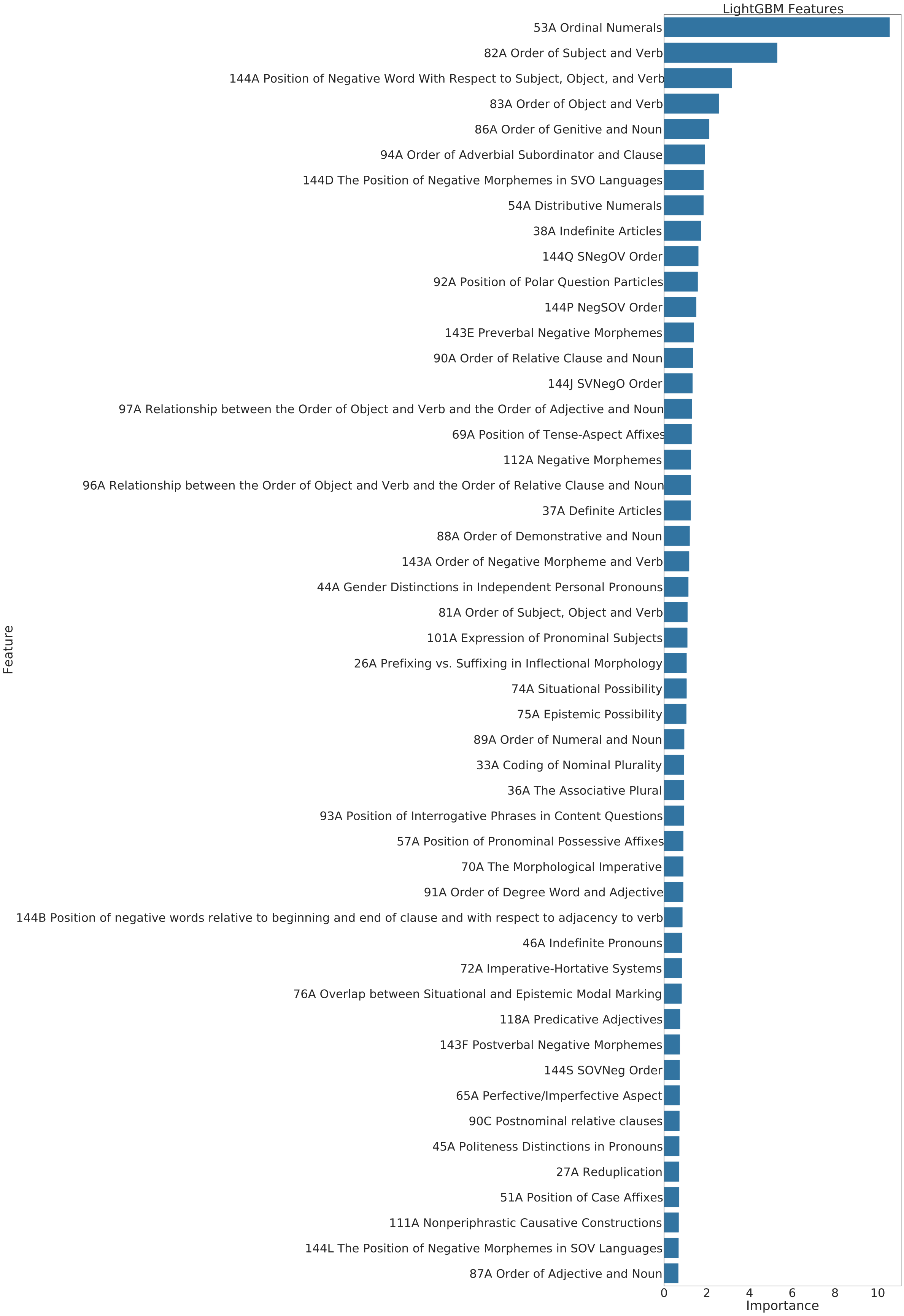
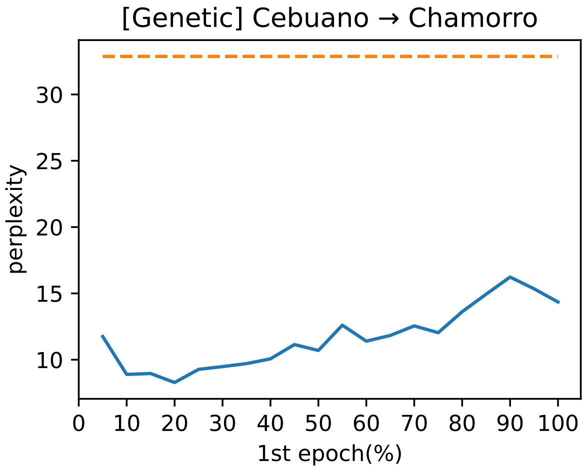
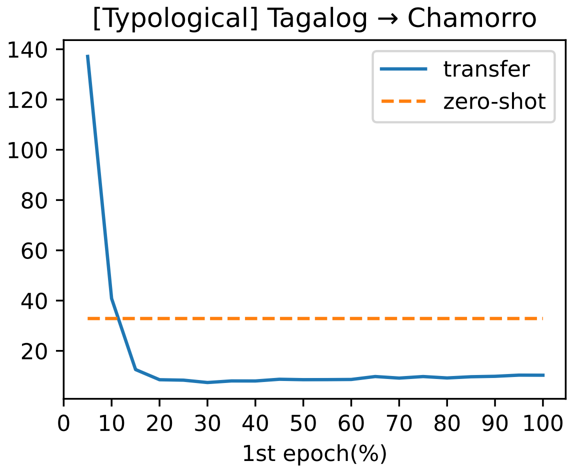
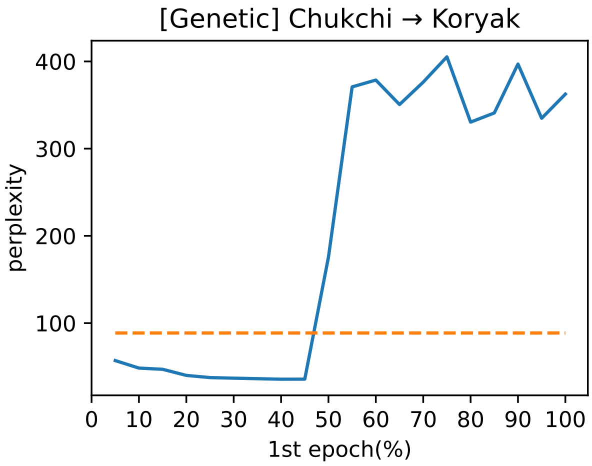


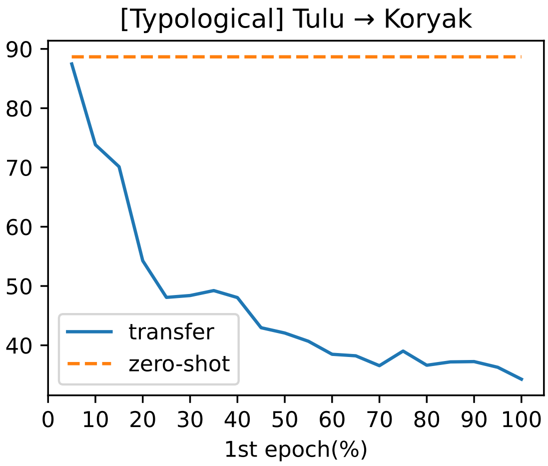
Figure 5: Feature importance measured with the LGBM classifier

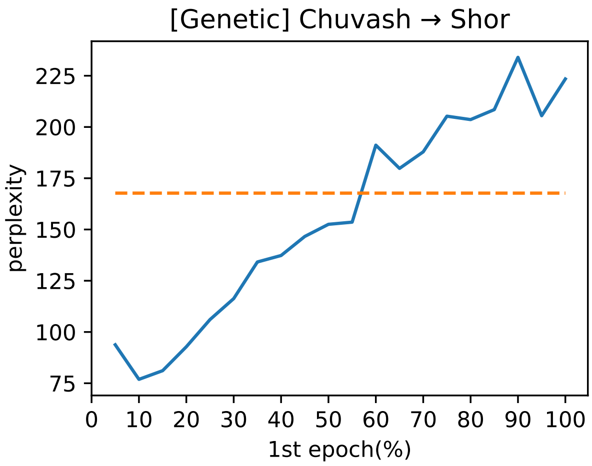
A.6 Appendix 6. Logs for both Genetic and Typological sources for every studied target language

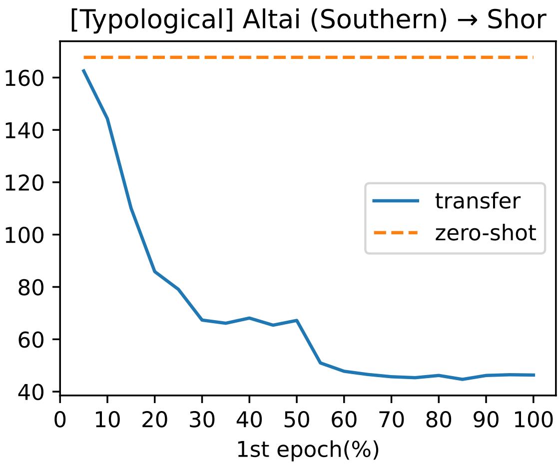


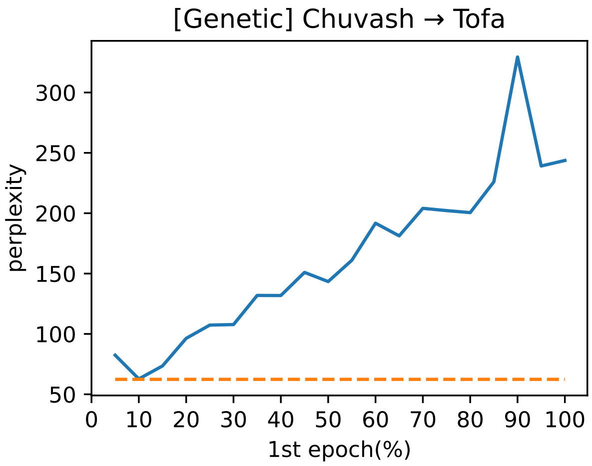


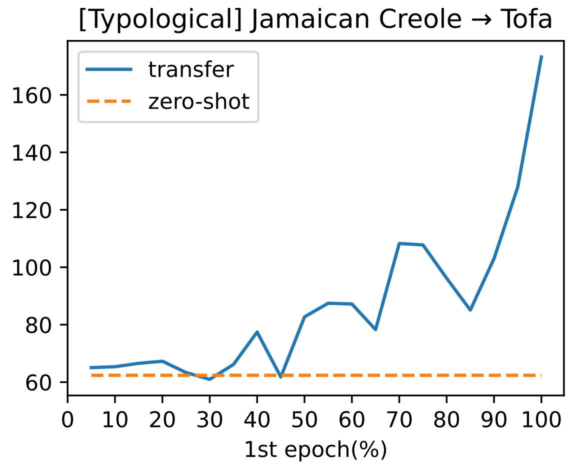


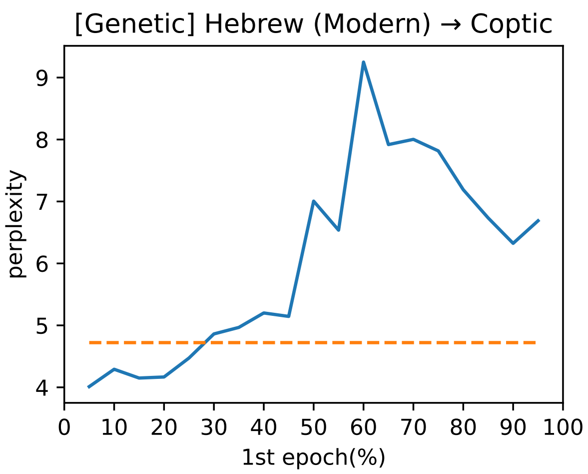


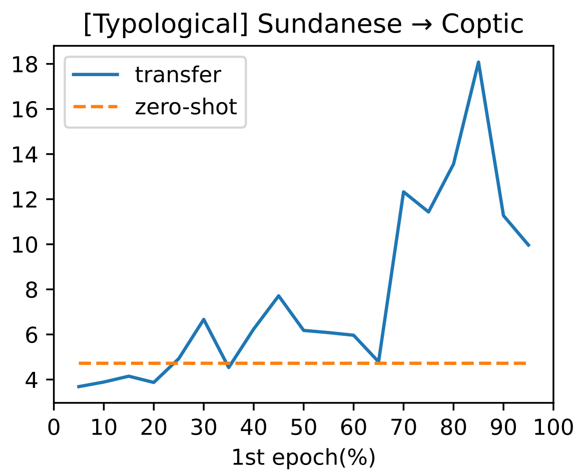


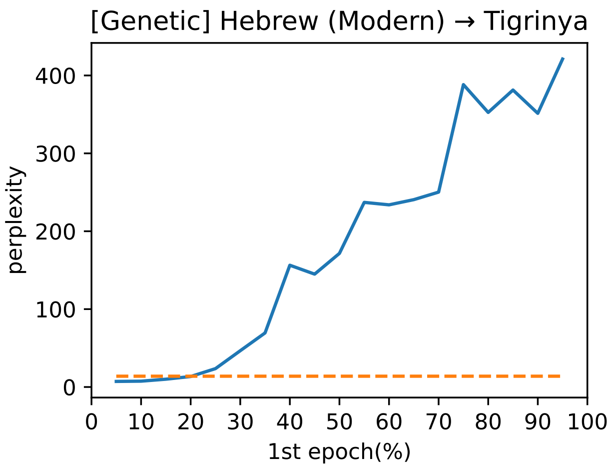


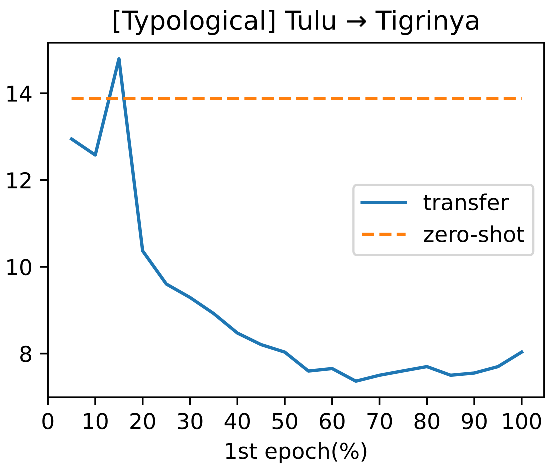

[Genetic] Chuvash → Tofa
perplexity
1st epoch(%)

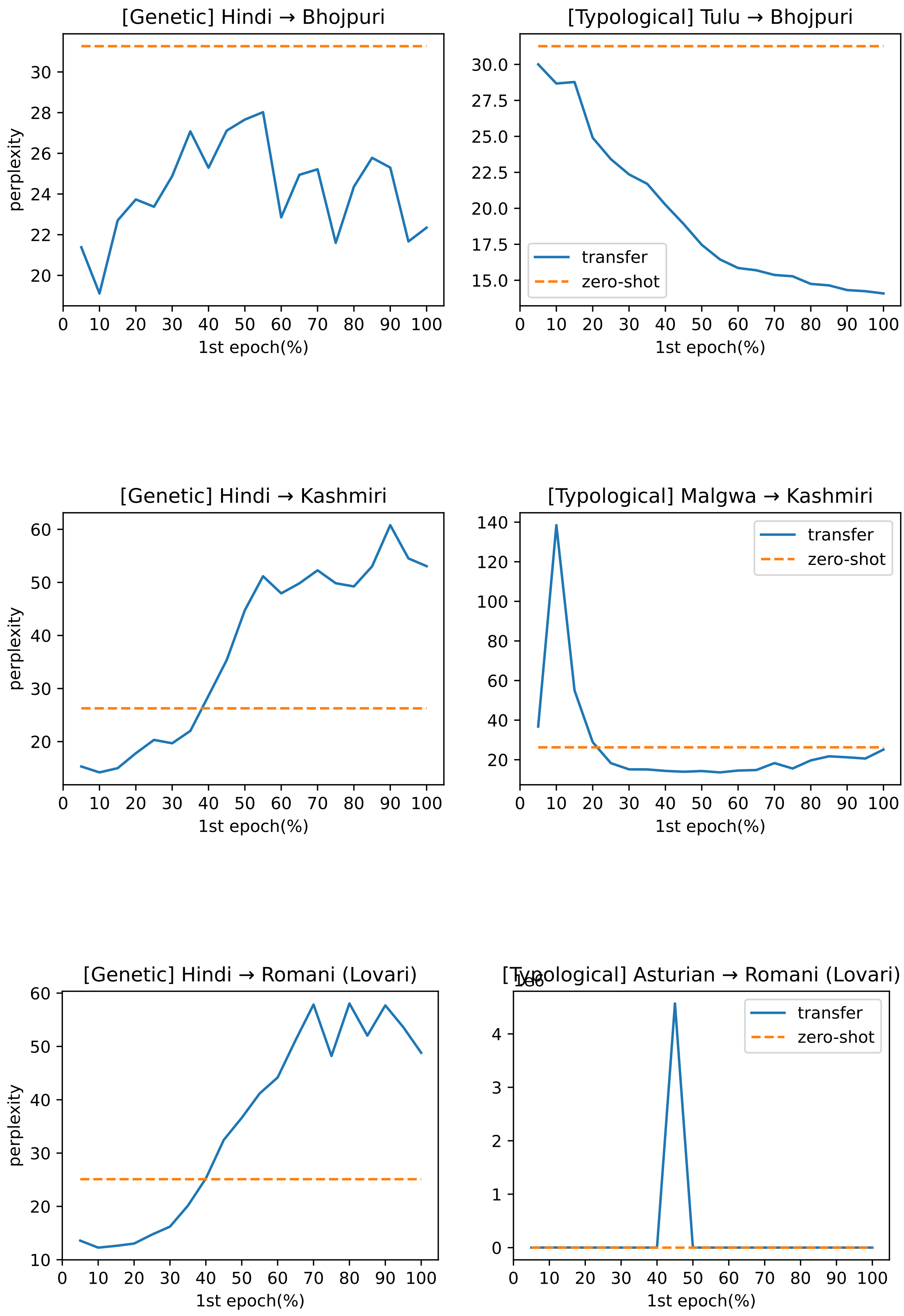

[Typological] Jamaican Creole → Tofa
transfer
zero-shot
1st epoch(%)


[Genetic] Hebrew (Modern) → Coptic
perplexity
1st epoch(%)

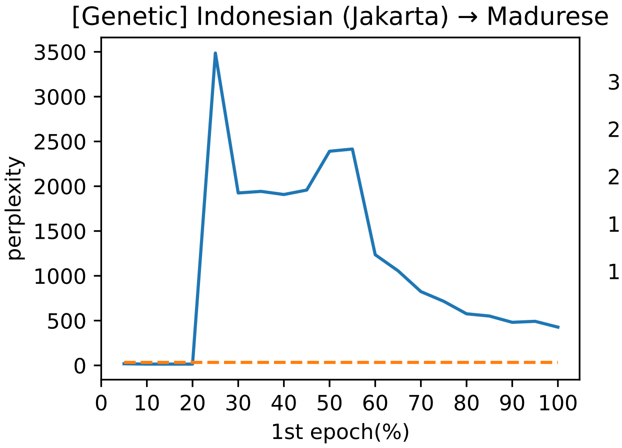

[Typological] Sundanese → Coptic
transfer
zero-shot
1st epoch(%)

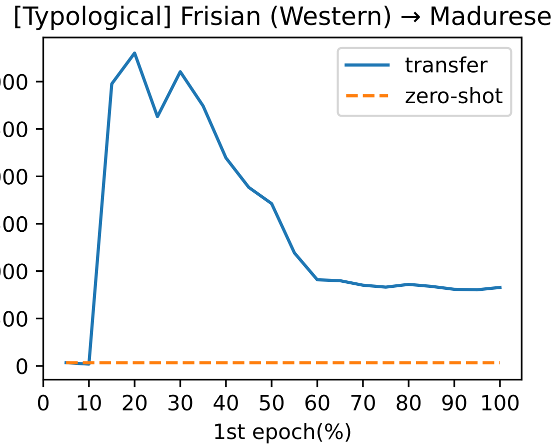

[Genetic] Hebrew (Modern) → Tigrinya
perplexity
1st epoch(%)

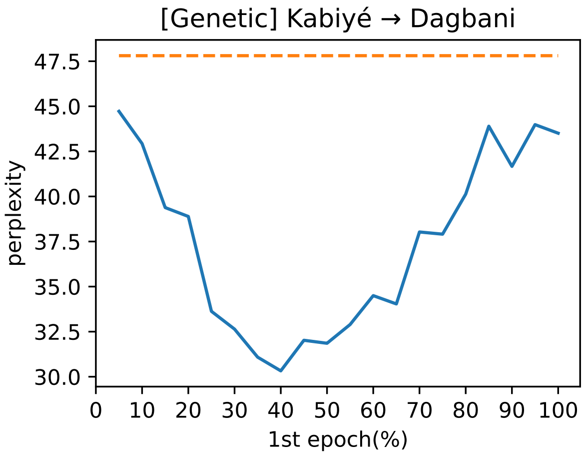

[Typological] Tulu → Tigrinya
transfer
zero-shot
1st epoch(%)

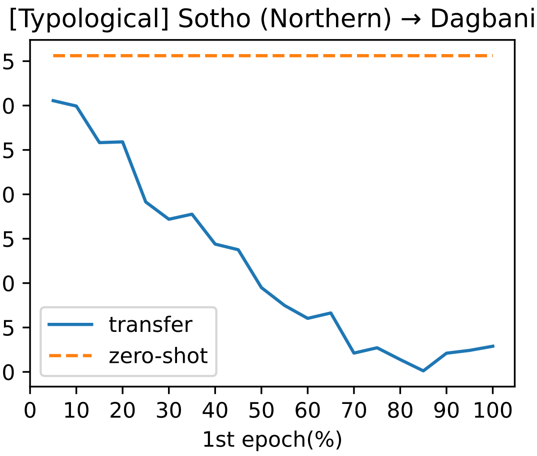


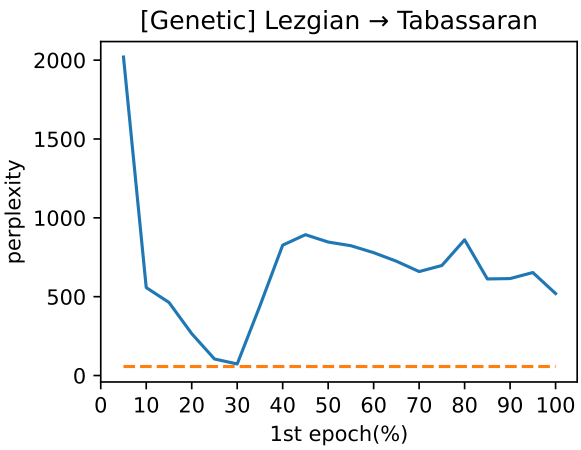
[Genetic] Hindi → Bhojpuri
[Typological] Tulu → Bhojpuri
[Genetic] Hindi → Kashmiri
[Typological] Malgwa → Kashmiri
[Genetic] Hindi → Romani (Lovari)
[Typological] Asturian → Romani (Lovari)
perplexity
1st epoch(%)
transfer
zero-shot
1e6

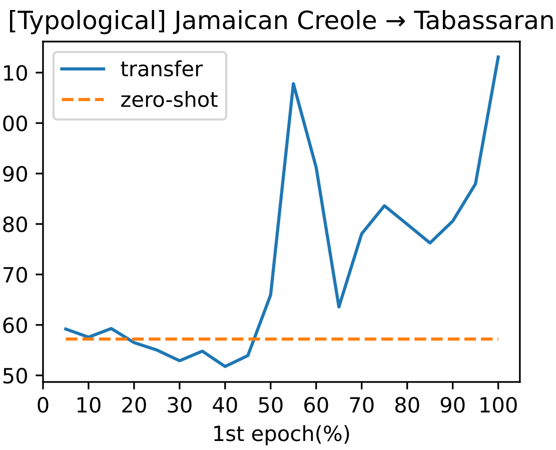

[Genetic] Indonesian (Jakarta) → Madurese
perplexity
1st epoch(%)

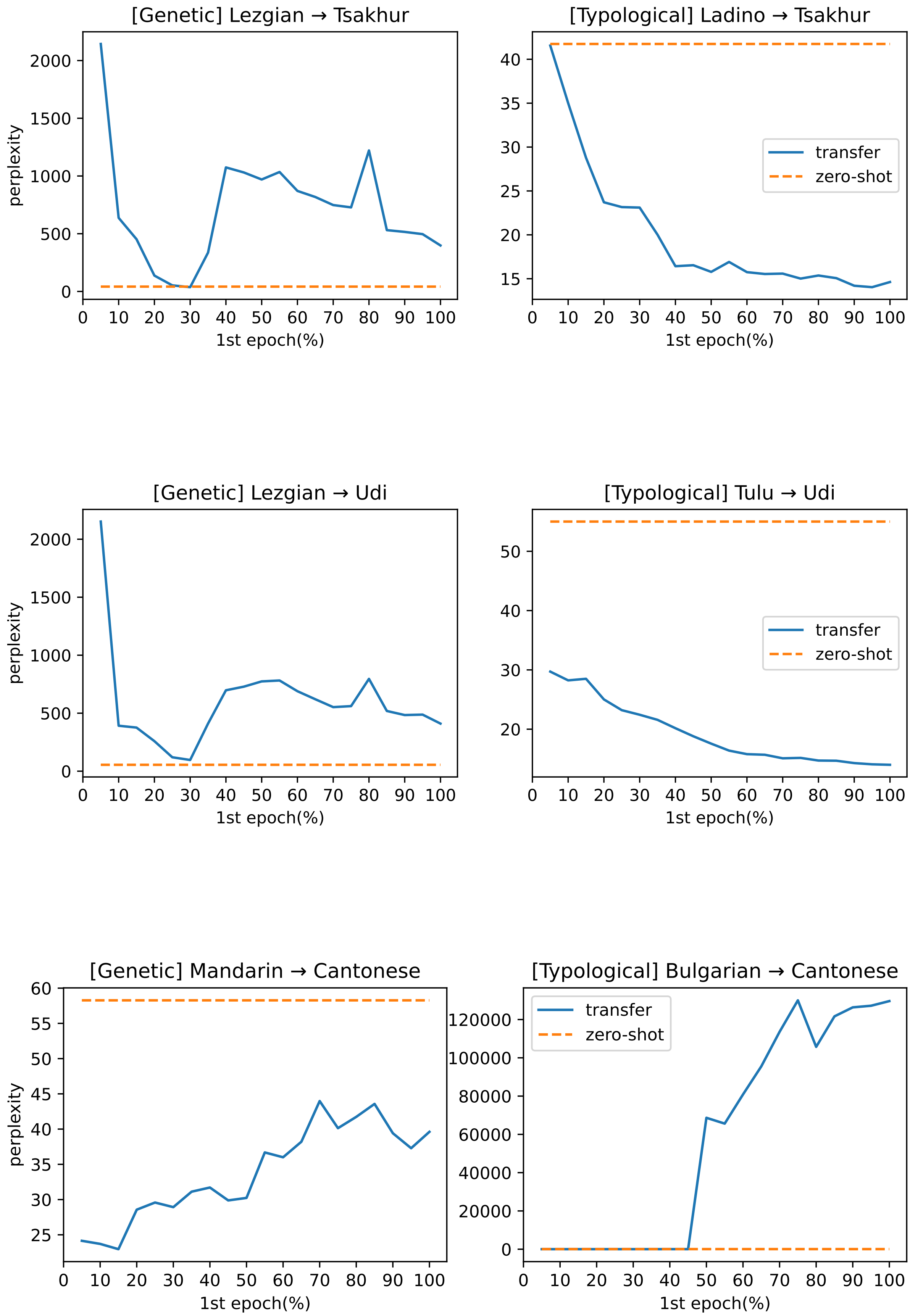

[Typological] Frisian (Western) → Madurese
transfer
zero-shot
1st epoch(%)


[Genetic] Kabiyé → Dagbani
perplexity
1st epoch(%)

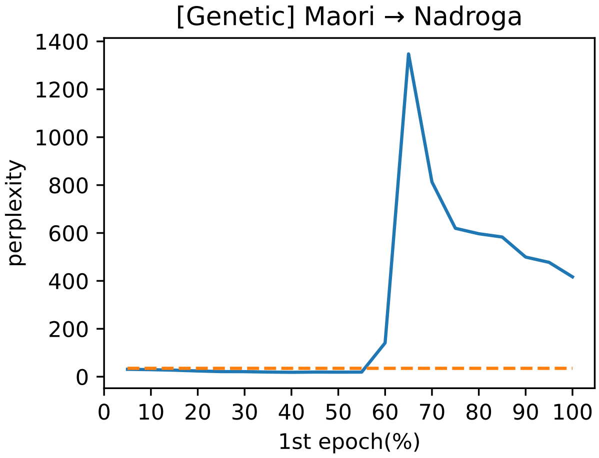

[Typological] Sotho (Northern) → Dagbani
transfer
zero-shot
1st epoch(%)

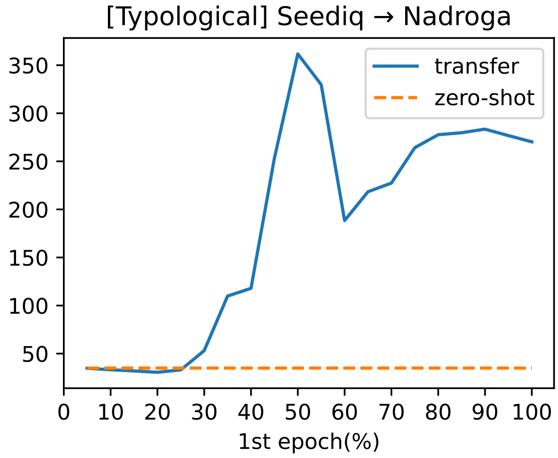

[Genetic] Lezgian → Tabassaran
perplexity
1st epoch(%)

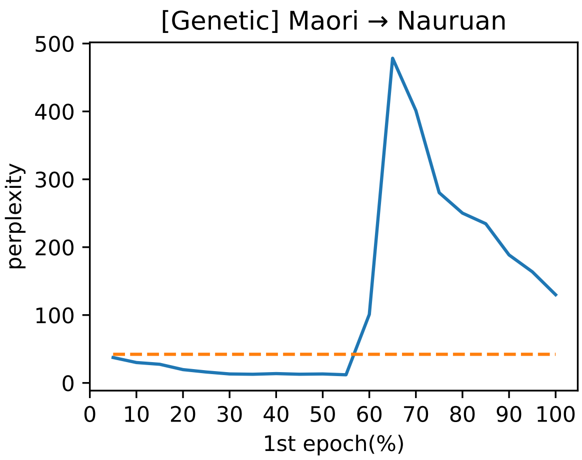

[Typological] Jamaican Creole → Tabassaran
transfer
zero-shot
1st epoch(%)

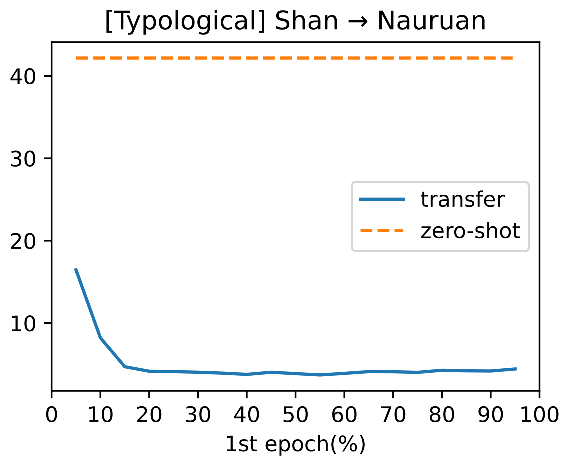


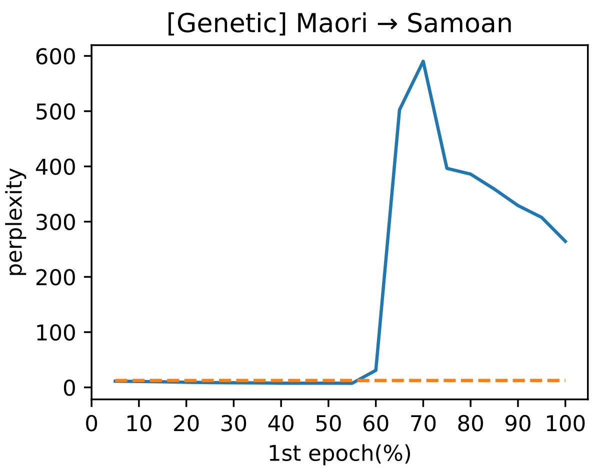
[Genetic] Lezgian → Tsakhur
perplexity
1st epoch(%)
[Typological] Ladino → Tsakhur
transfer
zero-shot
1st epoch(%)
[Genetic] Lezgian → Udi
perplexity
1st epoch(%)
[Typological] Tulu → Udi
transfer
zero-shot
1st epoch(%)
[Genetic] Mandarin → Cantonese
perplexity
1st epoch(%)
[Typological] Bulgarian → Cantonese
transfer
zero-shot
1st epoch(%)

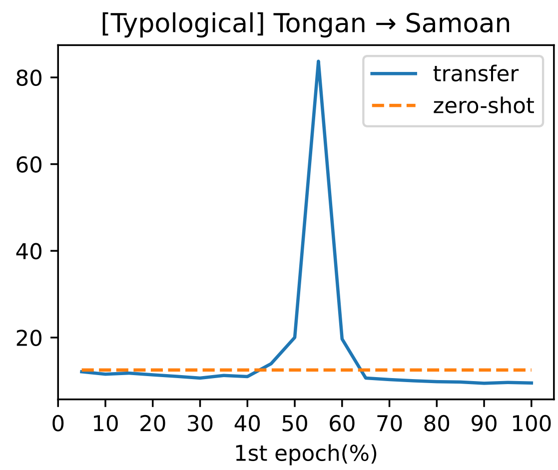

[Genetic] Maori → Nadroga
perplexity
1st epoch(%)

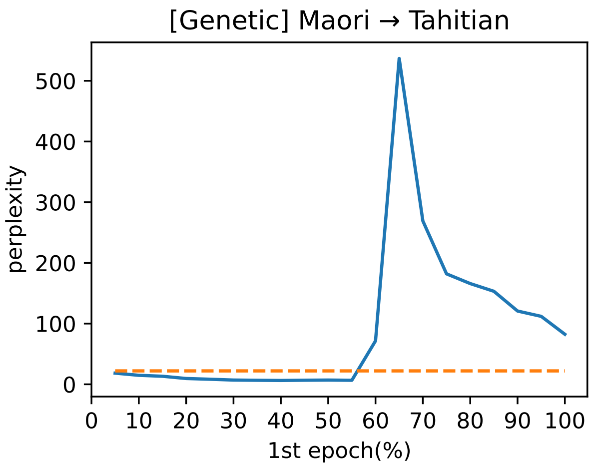

[Typological] Seediq → Nadroga
transfer
zero-shot
1st epoch(%)

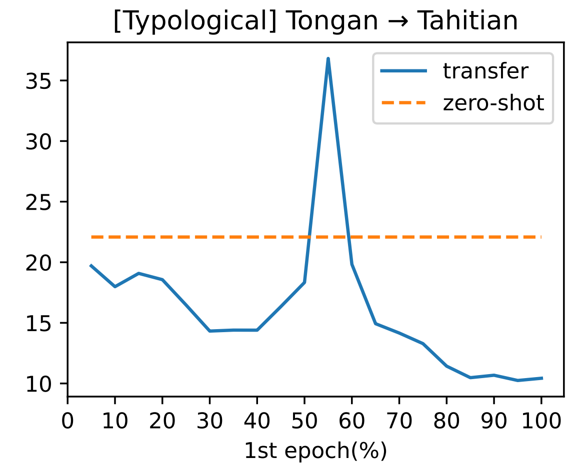

[Genetic] Maori → Nauruan
perplexity
1st epoch(%)

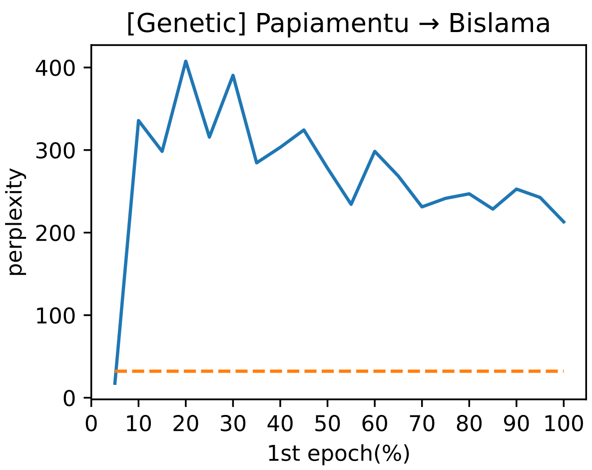

[Typological] Shan → Nauruan
transfer
zero-shot
1st epoch(%)

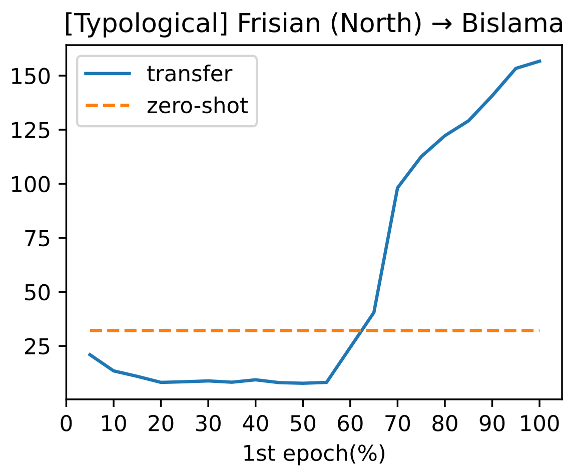

[Genetic] Maori → Samoan
perplexity
1st epoch(%)

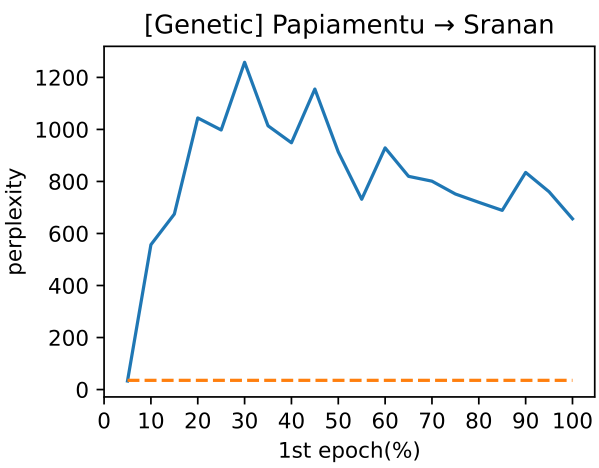

[Typological] Tongan → Samoan
transfer
zero-shot
1st epoch(%)

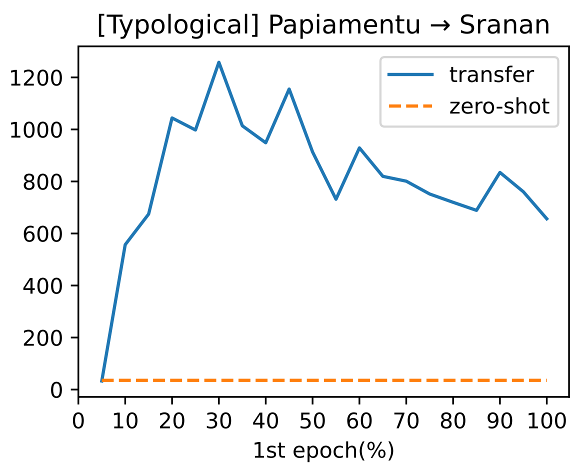

[Genetic] Maori → Tahitian
perplexity
1st epoch(%)

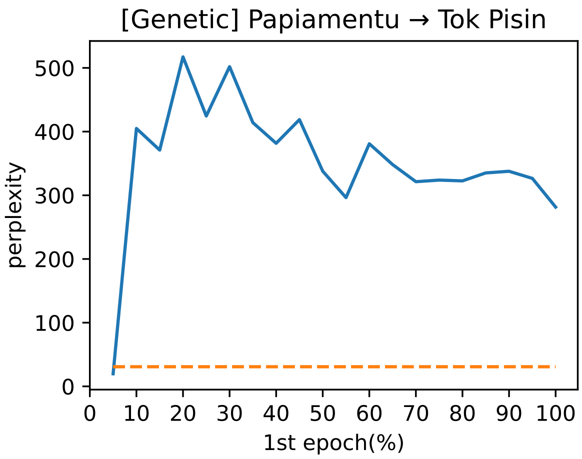

[Typological] Tongan → Tahitian
transfer
zero-shot
1st epoch(%)

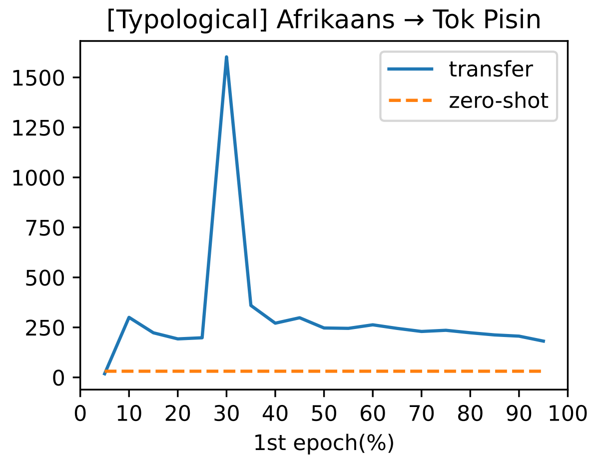

[Genetic] Papiamentu → Bislama
perplexity
1st epoch(%)

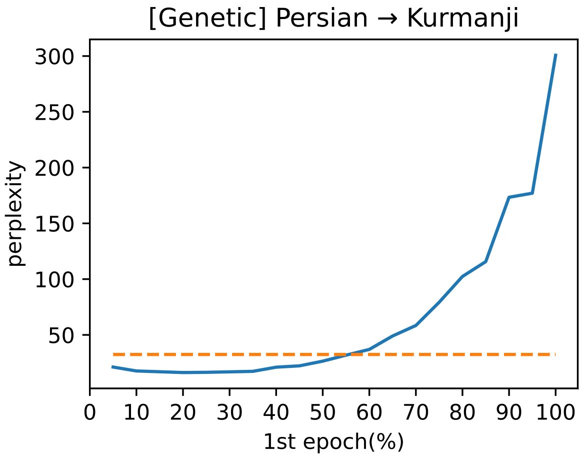

[Typological] Frisian (North) → Bislama
transfer
zero-shot
1st epoch(%)

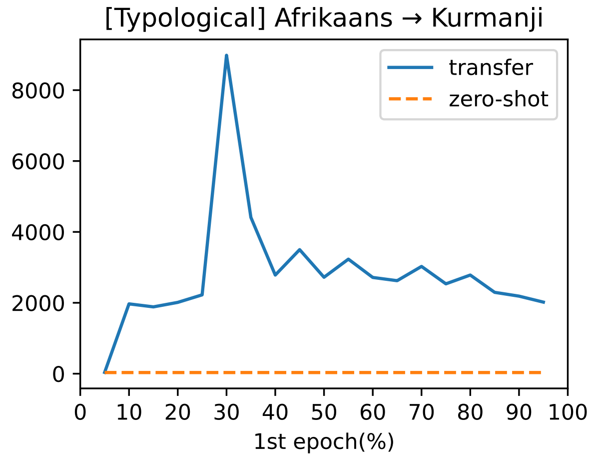

[Genetic] Papiamentu → Sranan
perplexity
1st epoch(%)

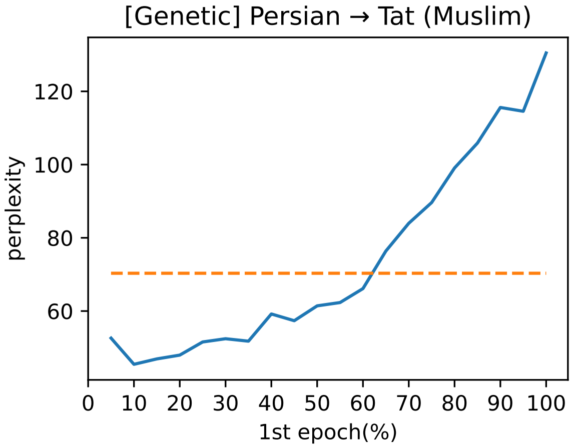

[Typological] Papiamentu → Sranan
transfer
zero-shot
1st epoch(%)

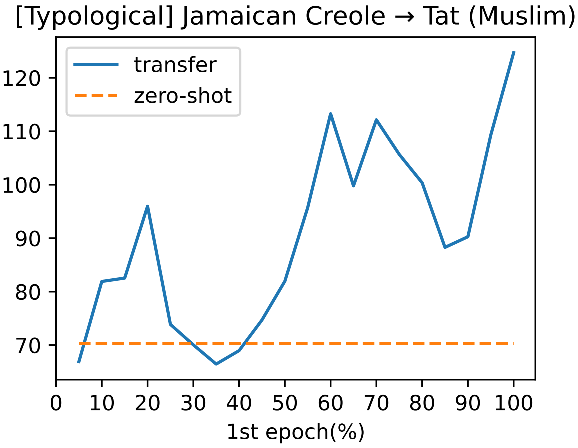

[Genetic] Papiamentu → Tok Pisin
perplexity
0
100
200
300
400
500
0 10 20 30 40 50 60 70 80 90 100
1st epoch(%)

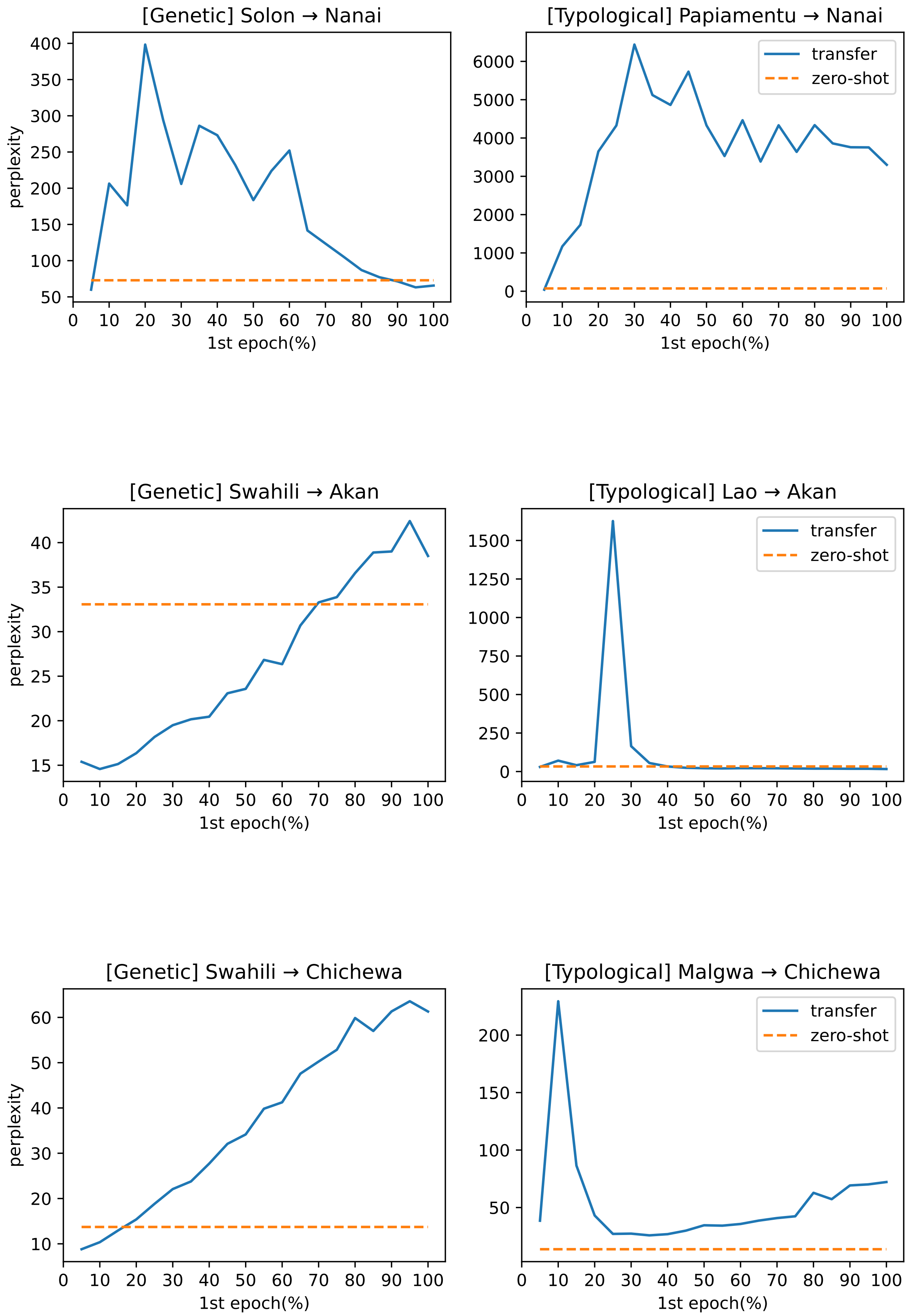

[Typological] Afrikaans → Tok Pisin
transfer
zero-shot
0
250
500
750
1000
1250
1500
0 10 20 30 40 50 60 70 80 90 100
1st epoch(%)


[Genetic] Persian → Kurmanji
perplexity
50
100
150
200
250
300
0 10 20 30 40 50 60 70 80 90 100
1st epoch(%)

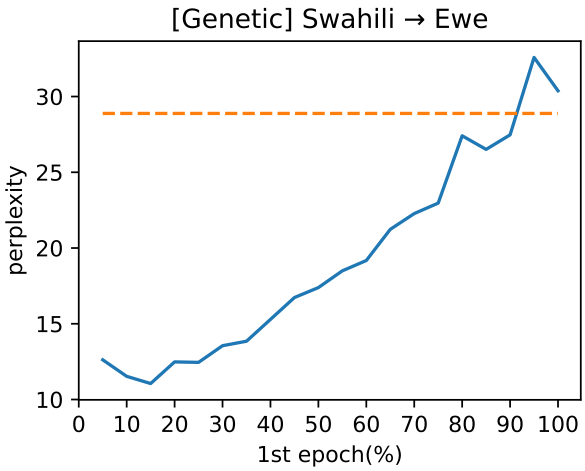

[Typological] Afrikaans → Kurmanji
transfer
zero-shot
0
2000
4000
6000
8000
0 10 20 30 40 50 60 70 80 90 100
1st epoch(%)

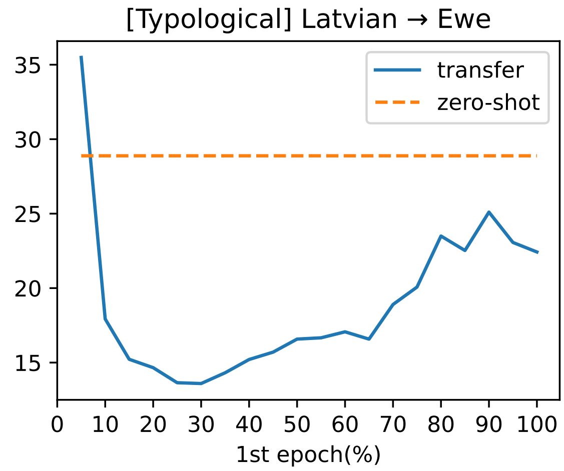

[Genetic] Persian → Tat (Muslim)
perplexity
60
80
100
120
0 10 20 30 40 50 60 70 80 90 100
1st epoch(%)

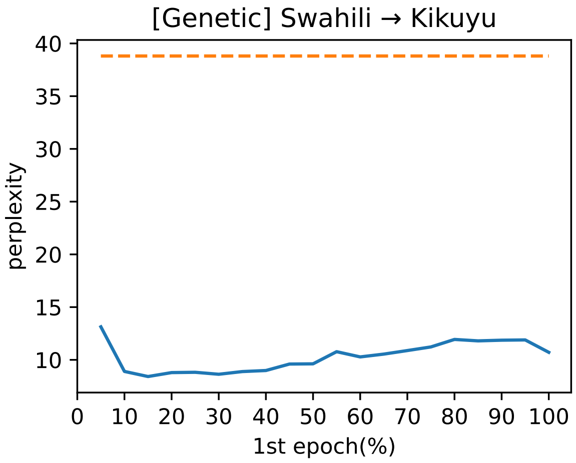

[Typological] Jamaican Creole → Tat (Muslim)
transfer
zero-shot
70
80
90
100
110
120
0 10 20 30 40 50 60 70 80 90 100
1st epoch(%)

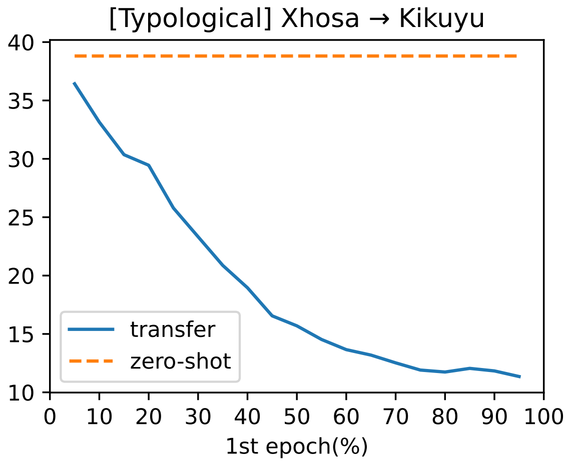


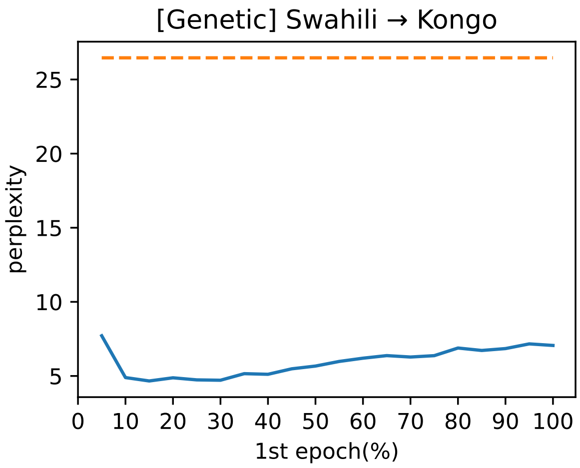
[Genetic] Solon → Nanai
[Typological] Papiamentu → Nanai
[Genetic] Swahili → Akan
[Typological] Lao → Akan
[Genetic] Swahili → Chichewa
[Typological] Malgwa → Chichewa
perplexity
1st epoch(%)
transfer
zero-shot

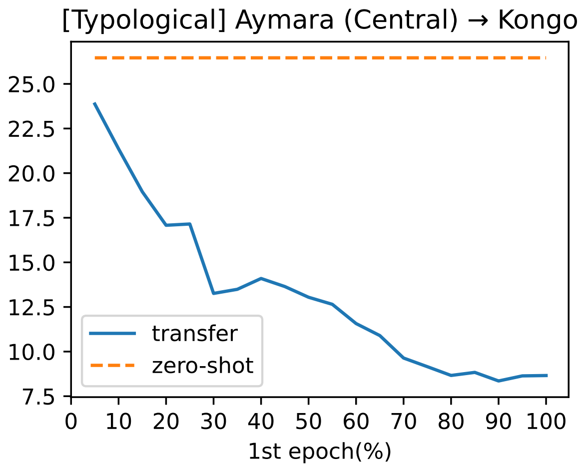

[Genetic] Swahili → Ewe
perplexity
1st epoch(%)
10 15 20 25 30
0 10 20 30 40 50 60 70 80 90 100

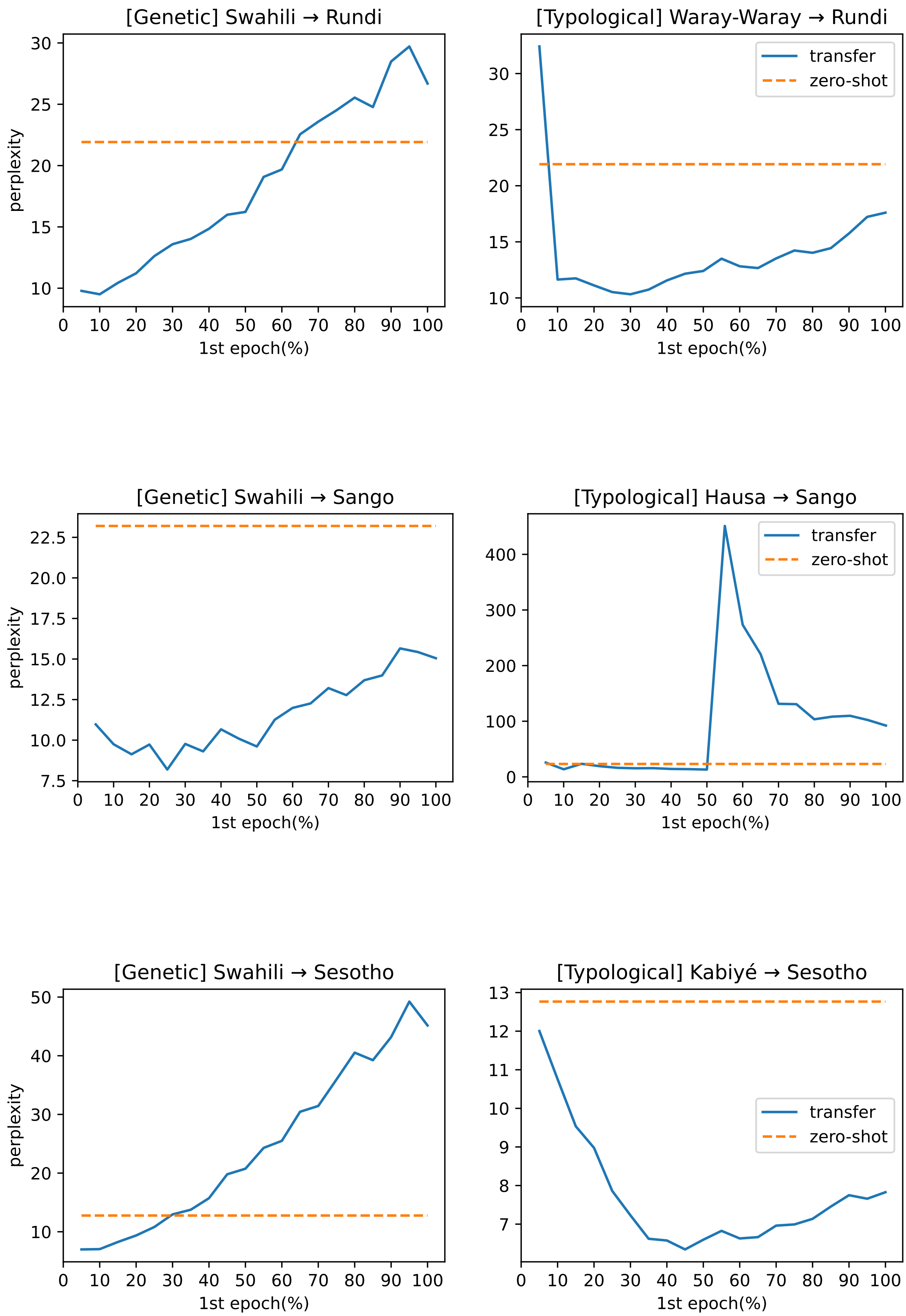

[Typological] Latvian → Ewe
transfer
zero-shot
15 20 25 30 35
0 10 20 30 40 50 60 70 80 90 100
1st epoch(%)

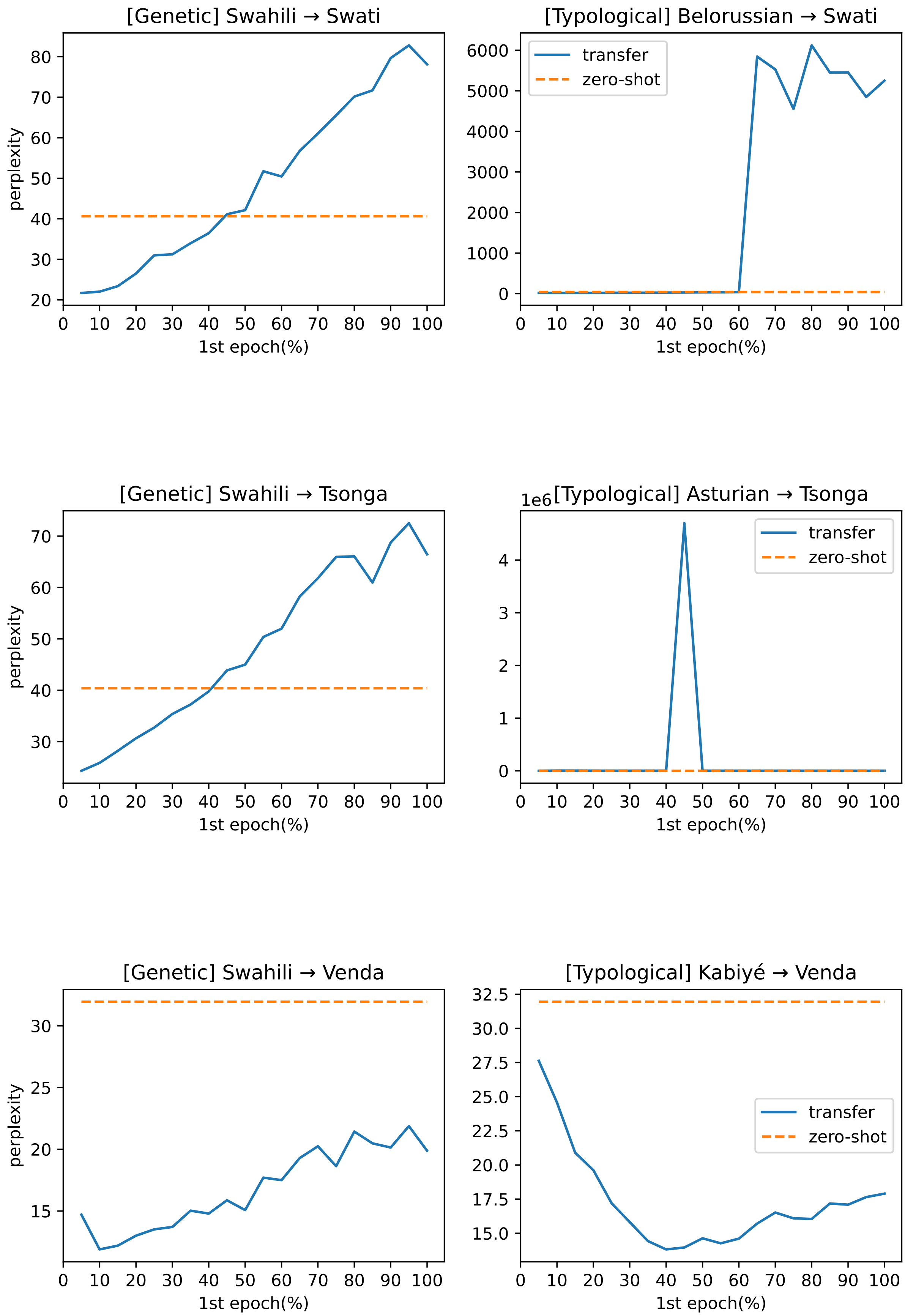

[Genetic] Swahili → Kikuyu
perplexity
10 15 20 25 30 35 40
0 10 20 30 40 50 60 70 80 90 100
1st epoch(%)

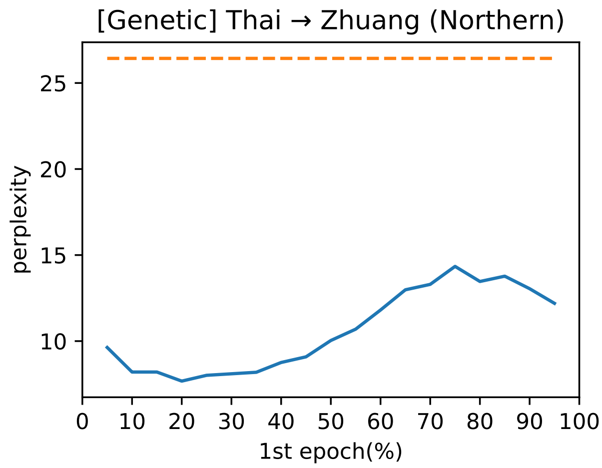

[Typological] Xhosa → Kikuyu
transfer
zero-shot
10 15 20 25 30 35 40
0 10 20 30 40 50 60 70 80 90 100
1st epoch(%)

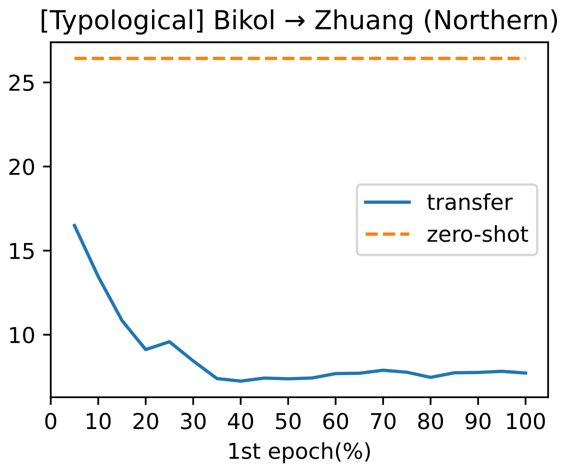

[Genetic] Swahili → Kongo
perplexity
5 10 15 20 25
0 10 20 30 40 50 60 70 80 90 100
1st epoch(%)

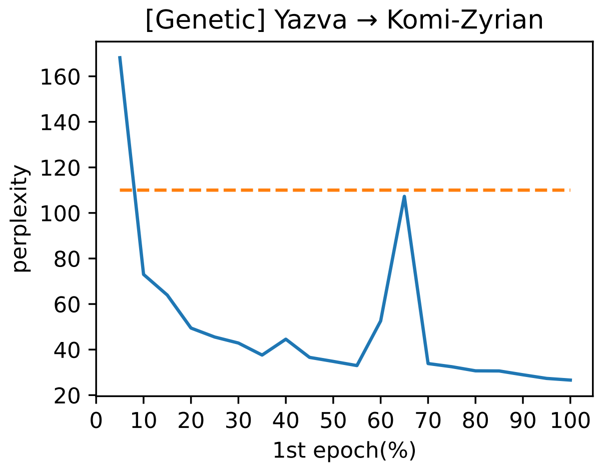

[Typological] Aymara (Central) → Kongo
transfer
zero-shot
7.5 10.0 12.5 15.0 17.5 20.0 22.5 25.0
0 10 20 30 40 50 60 70 80 90 100
1st epoch(%)

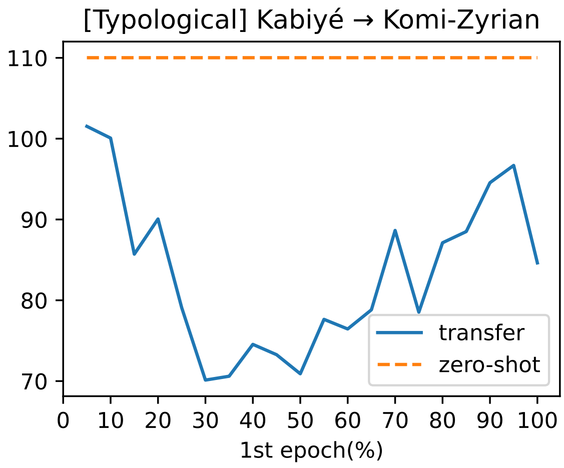

[Genetic] Swahili → Rundi
[Typological] Waray-Waray → Rundi
[Genetic] Swahili → Sango
[Typological] Hausa → Sango
[Genetic] Swahili → Sesotho
[Typological] Kabiyé → Sesotho
perplexity
1st epoch(%)
transfer
zero-shot


[Genetic] Swahili → Swati
[Typological] Belorussian → Swati
[Genetic] Swahili → Tsonga
1e6
[Typological] Asturian → Tsonga
[Genetic] Swahili → Venda
[Typological] Kabiyé → Venda
perplexity
1st epoch(%)
transfer
zero-shot


[Genetic] Thai → Zhuang (Northern)
perplexity
1st epoch(%)


[Typological] Bikol → Zhuang (Northern)
transfer
zero-shot
1st epoch(%)


[Genetic] Yazva → Komi-Zyrian
perplexity
1st epoch(%)


[Typological] Kabiyé → Komi-Zyrian
transfer
zero-shot
1st epoch(%)