XNLIeu: a dataset for cross-lingual NLI in Basque

Maite Heredia[1] **Julen Etxaniz**[1] **Muitze Zulaika**[2]
Xabier Saralegi[2] **Jeremy Barnes**[1] **Aitor Soroa**[1]
[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU
[2]Orai NLP Technologies
{maite.heredia}@ehu.eus

Abstract

XNLI is a popular Natural Language Inference (NLI) benchmark widely used to evaluate cross-lingual Natural Language Understanding (NLU) capabilities across languages. In this paper, we expand XNLI to include Basque, a low-resource language that can greatly benefit from transfer-learning approaches. The new dataset, dubbed XNLIeu, has been developed by first machine-translating the English XNLI corpus into Basque, followed by a manual post-edition step. We have conducted a series of experiments using mono- and multilingual LLMs to assess a) the effect of professional post-edition on the MT system; b) the best cross-lingual strategy for NLI in Basque; and c) whether the choice of the best cross-lingual strategy is influenced by the fact that the dataset is built by translation. The results show that post-edition is necessary and that the translate-train cross-lingual strategy obtains better results overall, although the gain is lower when tested in a dataset that has been built natively from scratch. Our code and datasets are publicly available under open licenses[1].

1 Introduction

The Natural Language Inference (NLI) task consists in classifying pairs of sentences –a premise and a hypothesis– according to their semantic relation: *entailment*, when the meaning of the premise entails that of the hypothesis; *contradiction*, when both sentences have opposing truth conditions and can not co-occur at the same time; and *neutral*, when both sentences are not semantically related (see Table 1 for examples).

NLI is an important task towards Natural Language Understanding (NLU), and is often used to test the semantic understanding of language models. It provides a general framework where different NLP tasks can be reframed, including information retrieval (Dušek et al., 2023), metaphor detection (Stowe et al., 2022) or relation extraction (Sainz et al., 2021). The NLI paradigm has also been proposed as a way to detect hallucination in Natural Language Generation (NLG) (Ji et al., 2023).

premise	Yesterday I saw an octopus at the beach.
entailment	I was at the beach yesterday.
contradiction	Yesterday I spent the whole day at home.
neutral	Octopi are my favourite animals.

Table 1: Example of a premise and three different hypotheses with the three possible relations.

XNLI (Conneau et al., 2018) is a popular benchmark widely used to evaluate cross-lingual NLI capabilities among languages. It comprises $7,500$ premise/hypothesis pairs in English that were manually translated to 14 high- and low-resource languages. In this paper we expand XNLI to include Basque, a low-resource language spoken in Spain and France (ISO-code: *eu*). The new dataset, dubbed XNLIeu, has been built by machine translating and post-editing the English XNLI. We release both the post-edited and machine-translated versions, which we used to assess to what extent professional post-edition is necessary to obtain a reliable NLI dataset.

Previous work has emphasized the importance of the origin of the train and test data in cross-lingual settings, i.e., whether they are original or created through translation. In particular, Artetxe et al. (2020) show that a mismatch in the origin between training and test data may have a serious impact on the results, particularly when comparing different cross-lingual strategies. Moreover, NLI datasets are known to be biased and contain artifacts that lead models to rely on superficial clues (Gururangan et al., 2018; Poliak et al., 2018; Tsuchiya, 2018; McCoy et al., 2019). To analyze the impact of these factors in XNLIeu, we have created a Native test set completely from scratch with original premises extracted from sources with content in Basque and

[1]https://github.com/hitz-zentroa/xnli-eu

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4177–4188

hypotheses provided by Basque speakers, which were specifically told to avoid such biases.

Using these datasets, we have conducted a series of experiments using mono- and multilingual language models for Basque, both discriminative and generative, and have tested different training variants for cross-lingual NLI in Basque. The experiments set a new baseline for NLI in Basque, and have served us to analyze the effect of professional post-edition compared to the automatic machine-translation system. We have also identified the most effective cross-lingual strategy for NLI in Basque, considering both translated and native sets.

This paper makes the following contributions:

- We develop and release a new dataset for cross-lingual NLI in Basque, which is created by translating the English XNLI, through machine-translation and post-edition. We also release a machine-translated only version of the dataset, as well as a small native dataset for comparison purposes.
- We conduct a series of cross-lingual Basque NLI experiments using several language models and following different cross-lingual strategies, and establish new baselines to facilitate research on Basque NLU.
- We provide a detailed analysis of the results of our experiments to assess the impact of using different models, strategies and data sources.

This paper is structured as follows: Section 2 covers some relevant research and resources related to the topic in hand, our dataset is further explained in Section 3, the description of the experiments and experimental settings in Section 4, the results are covered in Section 5, Section 6 includes the analysis of the errors in the outputs of our models, Section 7 a summary of the research and its conclusions; and there is a final section that expands on the limitations of our research.

2 Related work

Cross-lingual NLI. The best results on NLI benchmarks to date are based on supervised learning, which requires large amounts of training data that are only available for resource-rich languages such as English. Examples of English NLI datasets are the Stanford NLI corpus (Bowman et al., 2015), the Multi-genre NLI corpus (Williams et al., 2018) and the Adversarial NLI corpus (Nie et al., 2020). The NLI task is also included among the tasks of the popular NLU benchmarks GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019). Cross-lingual NLI is an alternative approach that leverages pre-trained multilingual models which are fine-tuned in resource-rich languages, then tested in the desired target language. This transfer approach, called *zero-shot*, is often compared to strategies that involve machine translation: *translate-train*, where the training set is translated to each target language and used to train the models on their respective language and *translate-test*, where the test set is translated to the high-resource language, usually English. Alternatively, large multilingual autoregressive models are also known to perform well in cross-lingual settings, by providing them with a set of correct input/label pairs as prompts for new inputs (Brown et al., 2020).

XNLI. The Cross-lingual NLI corpus (XNLI) (Conneau et al., 2018) comprises development and test sets in 15 high- and low-resource languages, meant as a cross-lingual benchmark for this task. Later, this corpus was expanded to include additional languages such as Korean (Ham et al., 2020).

NLI biases & artifacts. Most famous NLI datasets have also been reported to include biases and artifacts (Gururangan et al., 2018; Poliak et al., 2018; Tsuchiya, 2018; McCoy et al., 2019) that should be considered when analyzing the results, as they seem to have critical effects on the performance of systems. Artetxe et al. (2020) analyzes the effect that translated datasets have in cross-lingual settings, due to the so-called *translationese* (Volansky et al., 2013), and concludes that mismatches between the origin of training and evaluation datasets cause an important impact on the robustness of evaluation.

Evaluation of LLMs. Nowadays, the focus of the research on evaluation has shifted due to the outstanding growth of LLMs. These models can achieve comparable results to fine-tuned pre-trained models with zero-shot and few-shot approaches for evaluation. Consequently, the focus has shifted towards assessing the models' overall capabilities rather than their performance on specific tasks (Guo et al., 2023). However, low-resource languages like Basque lag behind in NLP development, and can still benefit considerably from semantic datasets for tasks like NLI, which

Label	Example
premise	*Dena idazten saiatu nintzen* 'I tried to write everything.'
entailment	*Nire helburua gauzak idaztea zen.* 'My goal was to write things'
contradiction	*Ez nintzen ezer idazten saiatu ere egin.* 'I didn't even try to write anything.'
neutral	*Aipatu zuen lan bakoitza idatzi nuen.* 'I wrote every paper he mentioned.'

Table 2: Examples from the XNLIeu dataset

was not previously available for this language.

3 The XNLIeu dataset

XNLIeu has been created by machine-translating the English XNLI development and test sets to Basque[2] followed by a manual post-edition step[3]. Some examples of XNLIeu are shown in Table 2. We also release the machine-translated version prior to post-edition, dubbed XNLIeu$_{\text{MT}}$, which we use to analyze the effect of post-edition (see Section 5.1).

Additionally, we created an original Basque test set from scratch, henceforth referred to as *native*, and compared the results with XNLIeu and XNLIeu$_{\text{MT}}$ (see Section 5.2). Inspired by Bowman et al. (2015) and Artetxe et al. (2020), we performed the following steps to build the native dataset:

- As a starting point, we extracted $5,000$ sentences from recent news in Basque, ensuring that they were not previously seen by the models used in the experiments. For this, we scraped Basque News sites and selected sentences from documents whose creation time was posterior to the release date of the pre-training corpora.

- From these initial sentences, we manually selected 207 sentences that we deemed appropriate for this task, and used them as premises. Examples of phrases that were discarded are headlines, image descriptions that do not include verbs, or questions, since it is not always possible to obtain the truth conditions of these types of sentences

- We redacted annotation guidelines that explain the task and provide examples to the annotators. In these guidelines, annotators are asked to be creative and to avoid as much as possible some of the annotation artifacts that have been found in the large datasets (Gururangan et al., 2018; Poliak et al., 2018), such as the use of negation to create contradictions. The detailed guidelines are described in Appendix C.

- With the assistance of native Basque speakers, one hypothesis was created per premise and label, resulting in three hypotheses per premise, with a total of 621 sentences.

- We performed a final series of minor corrections on the resulting dataset, correcting typos and ensuring that the meaning conveyed by the hypotheses entails the assigned label.

Finally, we also distribute a machine-translated version of the English MNLI training corpus to Basque, with a total of $392,702$ sentences, which we use in the translate-train experiments.

	XNLI (english)	XNLIeu	XNLIeu$_{\text{MT}}$	native
entailment	9.89	8.15	7.81	8.95
contradiction	10.39	8.73	8.39	9.94
neutral	11.4	9.31	8.98	9.41

Table 3: Average length of hypotheses for each semantic relation type in our three datasets, as well as the average for the original English instances.

3.1 Quantitative analysis

In this section we present a quantitative analysis of various aspects of the three developed datasets: XNLIeu, XNLIeu$_{\text{MT}}$ and the Native dataset.

Label distribution. Since there are three hypotheses for each premise in the dataset, the label distribution is perfectly balanced, resulting in no majority class and establishing the baseline accuracy at 33%. This applies to all three datasets.

Sentence length. The average token length for hypotheses for each semantic relation type, as shown in Table 3, indicates that there is a bias, as neutral hypotheses are longer on average, while entailed hypotheses tend to be shorter, likely because entailed sentences are often formed by omitting words from the premise (Gururangan et al.,

[2]All machine translations performed in the paper have been obtained using Elia at https://elia.eus/translator.

[3]We hired a professional translation service to perform the post-edition. As is customary, we asked for periodic samples of the post-editions to assert that the translation mistakes from the MT were being corrected and ensure the quality of the post-edited dataset.

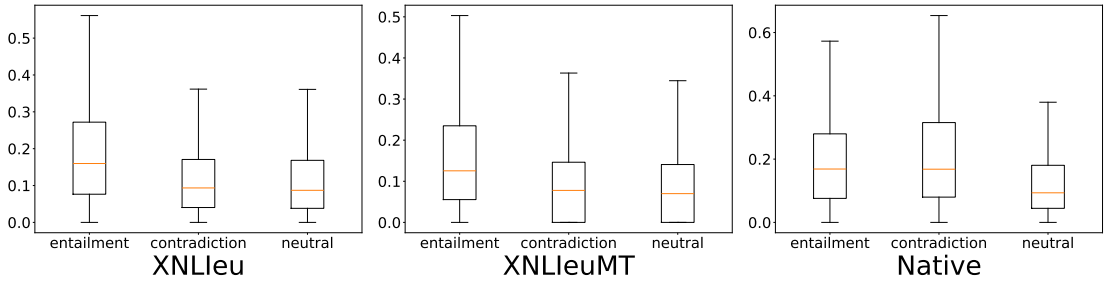


Figure 1: Box plots of the lexical overlap between premises and hypotheses calculated with cosine similarity of the three datasets.

2018). This bias is present in the original instances in English of the XNLI dataset and in XNLIeu and XNLIeu$_{\mathrm{MT}}$. The hypotheses of the Basque datasets tend to be shorter than the original English ones, but the unbalance between the different semantic relation types is maintained. The native set is also skewed, but in this case, the contradictions are slightly longer than neutral hypotheses, and entailments are still shorter on average.

Word frequency. Examining word frequency per label is insightful, especially since studies such as Gururangan et al. (2018) or Tsuchiya (2018) have reported that some NLI datasets exhibit a bias where the contradiction label is strongly associated with negation words. This seems to hold for the XNLIeu and XNLIeu$_{\mathrm{MT}}$ datasets. As we can see in Table 4, the word *ez* 'no' appears much more frequently in contradictions, and so do some other negations like *inork* 'nobody' or *inoiz* 'never'. It is plausible that models might be exploiting this feature as a form of shortcut learning for classification without even looking at the premise. The native dataset does not seem to be biased towards negation words, since the guidelines specifically asked the annotators to avoid using artifacts as much as possible (see Appendix C). It is interesting to note that among the most frequent words in this dataset, there are frequent references to the Basque culture: *euskaraz* 'in Basque', *euskara* / *euskal* 'Basque'or *Bilboko* 'from Bilbao'.

Lexical overlap. The lexical overlap between the premise and hypothesis has been calculated as the cosine similarity between the TF-IDF vector representations of both sentences. The results in Figure 1 show that in XNLIeu and XNLIeu$_{\mathrm{MT}}$ the highest overlap occurs between premises and entailed hypotheses. This is a known bias in NLI and is attributed to the fact that entailed hypotheses are easy to create by simply omitting parts of the premise (Gururangan et al., 2018). In contrast, this bias is not present in the native dataset, where on average the premises overlap mostly with both entailed and contradiction hypotheses, and less with neutral hypotheses.

	XNLIeu		XNLIeu$_{\mathrm{MT}}$		native	
entailment	no	0.58%	no	0.54%	in Basque	0.41%
	auxiliary[4]	0.24%	*auxiliary*	0.23%	film	0.24%
	something	0.19%	some	0.18%	*auxiliary*	0.24%
	some	0.18%	something	0.16%	movie	0.24%
	auxiliary	0.17%	like	0.13%	of the world	0.24%
contradiction	no	1.61%	no	1.65%	no	0.45%
	nobody	0.24%	nobody	0.23%	in Basque	0.34%
	never	0.2%	*auxiliary*	0.18%	Basque	0.28%
	auxiliary	0.18%	never	0.16%	my	0.23%
	my	0.16%	importance	0.14%	from Bilbao	0.23%
neutral	no	0.33%	no	0.31%	like	0.37%
	my	0.21%	dollar	0.2%	no	0.37%
	auxiliary	0.19%	my	0.2%	Basque	0.25%
	some	0.18%	*auxiliary*	0.16%	sometimes	0.25%
	like	0.15%	some	0.16%	people	0.25%

Table 4: Proportion of most frequent words of the three datasets, translated from Basque to English.

4 Experimental design

We have conducted a series of experiments on cross-lingual NLI for Basque, using different discriminative and generative language models, both mono- and multilingual. All models have been tested using the three datasets described in Section 3. We aim to determine if post-edition introduces significant changes to the dataset that enhance its reliability. We also want to compare the results on the XNLI-derived datasets with the native human-devised dataset, and analyze the effect of biases and artifacts introduced by translation. Since there is no training set in Basque for NLI, we consider different cross-lingual alternatives[5]:

- *Zero-Shot transfer*: We use multilingual discriminative models that have been pre-trained

[4]Auxiliaries are further discussed in Appendix A.

[5]The translate-test approach has not been implemented since the datasets have been originally translated from English to Basque, so back-translating them to English would not allow us to draw meaningful conclusions.

Discriminative		
Name	**Language**	**# of parameters**
IXAmBERT	Multilingual	177M
multilingual BERT	Multilingual	179M
XLM-RoBERTa (base)	Multilingual	279M
XLM-RoBERTa (large)	Multilingual	561M
BERTeus	Basque	124M
RoBERTa-eus Euscrawl	Basque	355M
Generative		
Latxa	Multilingual	7B
BLOOM	Multilingual	7.1B
XGLM	Multilingual	7.5B

Table 5: Details of the models used in the experiments.

at least in English and Basque. These models are then fine-tuned on the English MNLI corpus. In a further experiment, we explore fine-tuning with source languages beyond English.

- *Translate-train*: We machine-translate the English MNLI dataset to Basque, and use it to fine-tune the discriminative models (both multilingual and Basque monolingual).

- *Zero-shot prompting*: We directly test multilingual generative models that include Basque, without fine-tuning. We prompt the models by combining the premise and the hypothesis according to a template that is different for each possible label (See Appendix B).

Regarding the models, we have experimented with the following discriminative models: IXAmBERT (Otegi et al., 2020), multilingual BERT (Devlin et al., 2018), XLM-RoBERTa large (Conneau et al., 2019), BERTeus (Agerri et al., 2020) and RoBERTa-eus-large (Artetxe et al., 2022). Further details about these models can be found in Table 5. All of the models have been used in their cased version. For the BERT models, we have used a learning rate of $5e^{-5}$, and for the RoBERTa models, we have used a smaller learning rate of $10e^{-6}$, which is the only hyperparameter that has not been kept default, to avoid a degenerated solution. All models have been trained for 10 epochs, and the model selection has been performed on the development test. There has been no further attempt at hyperparameter optimization, since the goal was not to obtain the best possible model, but rather to compare the effects of the different sets and strategies. The models have been trained with three different random seeds to get the mean and the standard deviation and reduce the effects of randomness associated with initializing the weights and selecting the order of the training data. The code used for the experiments with discriminative models has been adapted from the code examples for fine-tuning for different tasks provided by Wolf et al. (2020).

zero-shot		
	XNLIeu	**XNLIeu$_{\text{MT}}$**
IXAmBERT	72.5 ($\pm 1.4e^{-3}$)	67.3 ($\pm 7.0e^{-3}$)
mBERT	60.1 ($\pm 5.7e^{-3}$)	57.9 ($\pm 1.2e^{-2}$)
XLM-RoBERTa base	73.4 ($\pm 3.5e^{-3}$)	69.0 ($\pm 9.0e^{-3}$)
XLM-RoBERTa large	**81.1** ($\pm 2.8e^{-3}$)	**75.4** ($\pm 2.0e^{-3}$)
translate-train		
	XNLIeu	**XNLIeu$_{\text{MT}}$**
IXAmBERT	75.9 ($\pm 6.4e^{-3}$)	71.3 ($\pm 4e^{-3}$)
mBERT	74.8 ($\pm 4.2e^{-3}$)	71.3 (± 0.0)
XLM-RoBERTa large	**83.8** ($\pm 6.0e^{-4}$)	**79.9** ($\pm 1.0e^{-3}$)
RoBERTa-euscrawl	83.0 ($\pm 7.1e^{-3}$)	78.6 ($\pm 2.0e^{-3}$)
BERTeus	79.0 ($\pm 4.2e^{-3}$)	74.9 ($\pm 8.0e^{-3}$)

Table 6: Accuracy of discriminative fine-tuned models tested with XNLIeu and XNLIeu$_{\text{MT}}$ datasets (mean and standard deviation of three runs). Best results in bold.

We have also tested three multilingual generative models that include Basque among their pretraining languages: BLOOM (BigScience Workshop et al., 2023), XGLM (Lin et al., 2022) and Latxa (Etxaniz et al., 2024), a model based on Llama 2 tuned for Basque with continual pretraining on Basque corpora. The prompts used in our experiments can be seen in Appendix B. As for evaluation, we select the label whose log-likelihood is highest, according to the model. The code used for testing the generative models is based on that included in the Language Model Evaluation Harness project (Gao et al., 2021).

Following usual practice, we use accuracy as our evaluation metric: the ratio of correctly classified instances divided by the total number of instances.

5 Results

In this section, we show the main results of our experiments and discuss the main findings. We start by analyzing the results on the datasets derived from XNLI (XNLIeu and XNLIeu$_{\text{MT}}$), followed by a comparison with those obtained using the native dataset. Finally, we detail the results of experiments that involved fine-tuning with source languages other than English.

5.1 Results for XNLIeu and XNLIeu$_{\text{MT}}$

The main results for the discriminative models can be seen in Table 6. All systems perform consistently better when evaluated on the post-

	XNLIeu	XNLIeu$_{\text{MT}}$
Latxa	**50.9**	**47.8**
BLOOM	49.5	47.5
XGLM	48.1	46.7

Table 7: Accuracy of generative models tested with XNLIeu and XNLIeu$_{\text{MT}}$ datasets using a zero-shot prompting approach. Best results in bold.

edited XNLIeu compared to the machine-translated XNLIeu$_{\text{MT}}$, and in some cases, the relative ranking among the models change, as is the case between multilingual BERT and IXAmBERT in the translate-test setting. Translate-train obtains better results overall on all models, and the difference is slightly higher in the XNLIeu$_{\text{MT}}$ dataset (7.3% accuracy points on average), where both training and test data have been created only through machine-translation. This result is consistent with the findings reported in Artetxe et al. (2020). Multilingual BERT is the model that improves the most with translate-train, probably because the presence of Basque at pre-training time was lower compared to the other models.

Table 7 shows the results obtained by the generative models. Once again, the models perform better when evaluated on the post-edited XNLIeu, but the performance gap is smaller compared with fine-tuned approaches. In any case, the results suggest that post-edition introduces significant changes to the dataset and is therefore important in order to obtain a reliable evaluation benchmark. We analyze this aspect further in Section 6.

5.2 Results for the native test set

Table 8 shows the results of the models when evaluated on the native dataset. The translate-train approach still yields better results than zero-shot transfer, but the difference in accuracy between both approaches is on average 2% percentage points smaller than those obtained with the translated sets. This is likely a consequence of the mismatch between the train and test sets, because in this setting, the former is built through translation text while the latter is natively written in Basque.

Discriminative models perform worse on the native dataset, with approximately 10% lower accuracy on average. While comparing results among different datasets is not always meaningful, we attribute the performance drop to the fact that the native dataset is less biased, as seen in Section 3.1. As a consequence, the models cannot rely on superficial patterns to deduce the relation between sentences, which makes this dataset especially challenging. Another possible cause is the notable presence of references to the Basque culture as it was sourced from original Basque materials.

zero-shot transfer	
IXAmBERT	64.0 ($\pm 9.0e^{-3}$)
mBERT	52.4 ($\pm 1.6e^{-2}$)
XLM-RoBERTa base	65.3 ($\pm 7.0e^{-3}$)
XLM-RoBERTa large	**73.8** ($\pm 7.0e^{-3}$)
translate-train	
BERTeus	68.4 ($\pm 1.0e^{-2}$)
IXAmBERT	65.6 ($\pm 1.0e^{-2}$)
mBERT	62.8 ($\pm 9.0e^{-3}$)
RoBERTa-euscrawl	75.2 ($\pm 7.0e^{-3}$)
XLM-RoBERTa large	**76.4** ($\pm 1.3e^{-2}$)
zero-shot prompting	
Latxa	**53.3**
BLOOM	49.8
XGLM	46.5

Table 8: Accuracy of discriminative (upper part) and generative (bottom part) models tested on the native dataset. Best results in bold.

Generative models yield results that are comparable to those obtained with machine-translated and post-edited sets. This result is a consequence of the zero-shot prompting strategy followed in generative models, which does not include fine-tuning, and therefore does not rely on examples that can induce bias in the model.

5.3 Choice of the source language

We have conducted additional typological experiments to test the impact of the choice of the source language in a zero-shot cross-lingual transfer setting for Basque. For this, we fine-tuned XLM-RoBERTa-base in 14 languages using machine-translated versions of the MNLI training data, as well as English, and tested them on XNLIeu, XNLIeu$_{\text{MT}}$ and the native test set. The results of these experiments are depicted in Table 9.

The table shows small differences in XNLIeu and XNLIeu$_{\text{MT}}$. We attribute these results to the fact that in this setting, both the training and test data come from translations, which lessens the importance of which source language to use. This is not the case for English, whose train data is original and not translated, but still it is not among the languages that achieve the highest results. When

	XNLIeu	XNLIeu$_{\mathrm{MT}}$	native
en	73.4 ($\pm 3.5e^{-3}$)	69.0 ($\pm 9.0e^{-3}$)	**65.3** ($\pm 7.0e^{-3}$)
ar	**73.9** ($\pm 2.6e^{-3}$)	**71.2** ($\pm 4.0e^{-3}$)	61.9 ($\pm 3.0e^{-3}$)
bg	73.2 ($\pm 8.9e^{-3}$)	<u>71.0</u> ($\pm 2.1e^{-3}$)	62.7 ($\pm 9.0e^{-3}$)
de	<u>73.9</u> ($\pm 5.3e^{-3}$)	70.4 ($\pm 7.0e^{-4}$)	63.5 ($\pm 8.0e^{-3}$)
el	73.7 ($\pm 1.7e^{-3}$)	70.7 ($\pm 7.0e^{-4}$)	63.6 ($\pm 7.0e^{-3}$)
es	73.7 ($\pm 5.2e^{-3}$)	70.3 ($\pm 7.0e^{-4}$)	<u>65.0</u> ($\pm 7.0e^{-3}$)
fr	73.7 ($\pm 4.9e^{-3}$)	69.9 ($\pm 7.1e^{-3}$)	63.3 ($\pm 2.1e^{-2}$)
hi	73.3 ($\pm 7.0e^{-3}$)	70.7 ($\pm 4.2e^{-3}$)	62.3 ($\pm 5.0e^{-3}$)
ru	72.9 ($\pm 1.5e^{-3}$)	69.7 ($\pm 2.1e^{-3}$)	62.2 ($\pm 6.0e^{-3}$)
sw	71.8 ($\pm 3.1e^{-3}$)	68.3 ($\pm 7.1e^{-3}$)	63.1 ($\pm 6.0e^{-3}$)
th	73.0 ($\pm 6.7e^{-3}$)	70.2 ($\pm 4.2e^{-3}$)	64.1 ($\pm 6.0e^{-3}$)
tr	73.5 ($\pm 6.2e^{-3}$)	70.9 ($\pm 7.0e^{-4}$)	63.6 ($\pm 7.0e^{-3}$)
ur	66.5 ($\pm 4.6e^{-3}$)	65.0 ($\pm 1.4e^{-3}$)	56.0 ($\pm 1.1e^{-2}$)
vi	72.6 ($\pm 1.1e^{-2}$)	69.6 ($\pm 7.8e^{-3}$)	62.4 ($\pm 1.5e^{-2}$)
zh	71.8 ($\pm 7.0e^{-3}$)	69.7 ($\pm 2.1e^{-3}$)	62.0 ($\pm 6.0e^{-3}$)

Table 9: Zero-shot cross-lingual transfer accuracy of XLMRoBERTa fine-tuned in different languages (mean and standard deviation of three runs). Best results in bold, second best underlined.

tested on the native dataset factors such as proximity between languages and loanword frequency gain relevance, as shown in the table, and the difference among languages is higher. Choosing English or Spanish yields similar results, while the performance when any other language is selected is noticeably lower.

6 Analysis

This Section provides additional analyses of the results. We begin by considering the performance of the best model on a per-label basis, followed by a manual comparison of the model outputs on the XNLIeu and XNLIeu$_{\mathrm{TM}}$ datasets to analyze the effects of post-edition.

6.1 Results per label

Figure 2 shows the confusion matrices on each label (entailment/neutral/contradiction) corresponding to the model and setting that performed best, XLM-RoBERTa large fine-tuned in Basque. For both XNLIeu and XNLIeu$_{\mathrm{MT}}$, the label that gets the higher F1 score is contradiction (87.7 and 83.4 respectively), followed by entailment (83 and 79.1), while neutral instances obtain the worst F1 score overall (80.7 and 76.4). This is in accordance with the analysis performed in Section 3.1, which indicates the presence of biases in these datasets, as well as in the training dataset. The results suggest that the models do rely on those biases, for instance by classifying instances where the hypothesis contains negative words as contradictions, or those where the hypothesis is short and has large lexical similarity with premises as entailment. On the other hand, no specific biases were detected in neutral instances, and consequently, it is more difficult for models to correctly classify them.

Section 3.1 reveals that the native dataset does not suffer from such apparent biases, and this is again reflected in the results depicted in Figure 2 for this dataset (right part). While contradiction is still the label with the best F1 score (80.3), now the label that attains the worst F1 is entailment (71.2), and the second-best is neutral (73.9).

6.2 Effects of post-edition

Section 5 reveals that systems perform consistently worse when evaluated on the machine-translated XNLIeu$_{\mathrm{MT}}$ dataset compared to the post-edited XNLIeu. So as to get a deeper insight into this result, we performed an analysis on XNLIeu and XNLIeu$_{\mathrm{MT}}$ by selecting instances that have been correctly predicted in one dataset and wrongly predicted in the other. The analysis reveals that XNLIeu$_{\mathrm{MT}}$ often contains translation errors that change the relation between premise and hypothesis, and that when post-editing the professional translators corrected those errors. The most frequent error converts entailment and contradiction hypotheses to neutral. Common translation errors include:

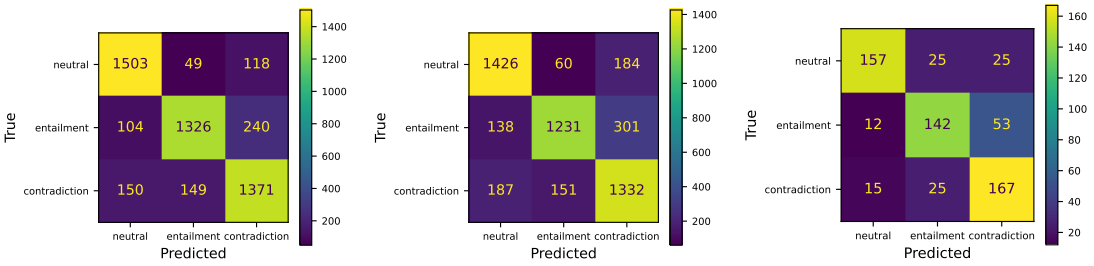
- Changing the polarity of a sentence from negative to positive or vice versa.

(1) Original: No, I live off campus.
MT: *ez naiz campusetik kanpo bizi*
'I don't live off campus'

- Using an incorrect auxiliary verb, which can have a detrimental effect and completely change the meaning of a sentence.

(2) Original: I was still scared.
MT: *eta oraindik beldurra ematen dit*
'I am still scared'

- Omitting crucial information from the original sentence or occasionally creating nonsensical sentences.



XNLIeu XNLIeu$_{\text{MT}}$ Native

Figure 2: Confusion matrices for the XLM-RoBERTa large fine-tuned in Basque, our best model, tested in our three datasets. Best viewed in color.

(3) Original: I like feeling myself.
MT: *Nik neuk gustuko dut ontzia.*
'I like the vessel myself'

On the other hand, there do not seem to be clear patterns in those instances that have been correctly predicted on XNLIeu$_{\text{MT}}$ and incorrectly on XNLIeu. We have only found a handful of examples where the original label of XNLI is ambiguous and post-edition introduces necessary changes to make the translations accurate and fluent, which can alter the relation between both sentences.

7 Conclusions

In this work, we introduce XNLIeu, a new dataset for cross-lingual NLI in Basque. XNLIeu is developed by machine-translating the English part of XNLI followed by a post-edition step with the assistance of professional translators. Along with XNLIeu we release the full machine-translated version, as well as a Basque native version carefully built to avoid known biases in NLI datasets. We have conducted a series of cross-lingual Basque NLI experiments using a set of language models and different cross-lingual strategies. The experiments show that translate-train is the best strategy, particularly when there is no mismatch between the origin of the train and test data. In the native dataset, translate-train still yields the best results, but the difference is comparatively smaller. This finding aligns with prior research examining the effects of translation-based datasets. We also manually analyze the results of the models and find that machine-translation often introduces artifacts that change the meaning of the premises or hypotheses, and that professional translators correct those errors when post-editing. We conclude that post-edition is a crucial step towards reliable evaluation of cross-lingual NLI.

All of the datasets developed in this paper are publicly available under the same licenses as XNLI. We believe that they are an important resource that will contribute to filling the gaps in resources that exist in Basque, which can hinder the development of research and applications with a focus on semantics in this language.

Limitations

Some limitations to this study should be taken into account, specially in the design of future research.

We have centered our work around the Basque language, which is considered to be a low-resource language. This means that, although some LLMs feature Basque in their training, there is not as much data and tools available as for other languages like English or Spanish. This was the main motivation for this research, but there is no prior work about NLI in Basque to be used as a reference, specifically in the experimental design and the interpretation of the results of the experiments.

Generative models are becoming more complex and versatile and are currently a popular subject of investigation. Most modern evaluation approaches are not focused on creating large corpora for specific tasks, but rather on testing generative models using prompt engineering and zero-shot or few-shot strategies. Our approach may seem outdated, as our research has focused mainly on the creation of our datasets and discriminative models, and generative models have only been tested with a zero-shot prompting approach. Future research for NLI in Basque should extend this line of research to account for the most recent developments and should include more insight into effective prompts and experiments performed with strategies other than zero-shot. However, we believe that the creation of our dataset and the approach we have followed

are still pertinent for a low-resource language like Basque, which unfortunately does not include all the necessary resources to fully leverage the most recent advances brought by generative models, and can take advantage of a task like NLI, which enables the development of semantic applications and is useful for transfer-learning into a lot of different tasks.

Acknowledgements

This work has been partially supported by the Basque Government (Research group funding IT-1805-22 and ICL4LANG project, grant no. KK-2023/00094) as well as the DeepR3 project (TED2021-130295B-C31) founded by MCIN/AEI/10.13039/501100011033 and European Union NextGeneration EU/PRTR. Julen Etxaniz holds a PhD grant from the Basque Government (PRE_2023_2_0060).

References

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: the case for basque.

Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez-de Viñaspre, and Aitor Soroa. 2022. Does corpus quality really matter for low-resource languages? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7383–7390, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Roman Dušek, Aleksander Wawer, Christopher Galias, and Lidia Wojciechowska. 2023. Improving domain-specific retrieval by nli fine-tuning.

Julen Etxaniz, Oscar Sainz, Naiara Perez Miguel, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for basque. *arXiv preprint*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating large language models: A comprehensive survey.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 422–430, Online. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual language models.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Arantxa Otegi, Aitor Agirre, Jon Ander Campos, Aitor Soroa, and Eneko Agirre. 2020. Conversational question answering in low resource scenarios: A dataset and case study for basque. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 436–442.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. Label verbalization and entailment for effective zero and few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Vered Volansky, Noam Ordan, and Shuly Wintner. 2013. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Most frequent words in original Basque

Table 10 shows the original words that have been translated to English in Table 4.

	XNLIeu		XNLIeu$_{MT}$		native	
	ez	0.58%	ez	0.54%	euskaraz	0.41%
	nuen	0.24%	nuen	0.23%	filma	0.24%
entailment	zerbait	0.19%	batzuek	0.18%	dezakezu	0.24%
	batzuek	0.18%	zerbait	0.16%	pelikula	0.24%
	daitezke	0.17%	gustatzen	0.13%	munduko	0.24%
	ez	1.61%	ez	1.65%	ez	0.45%
	inork	0.24%	inork	0.23%	euskaraz	0.34%
contradiction	inoiz	0.2%	nuen	0.18%	euskara	0.28%
	nuen	0.18%	inoiz	0.16%	nire	0.23%
	nire	0.16%	axola	0.14%	bilboko	0.23%
	ez	0.33%	ez	0.31%	gustatzen	0.37%
	nire	0.21%	dolar	0.2%	ez	0.37%
neutral	nuen	0.19%	nire	0.2%	euskal	0.25%
	batzuek	0.18%	nuen	0.16%	batzuetan	0.25%
	gustatzen	0.15%	batzuek	0.16%	jende	0.25%

Table 10: Proportion of most frequent words in Basque.

Some common words (*nuen, daitezke, dezakezu*) have been translated to English as *auxiliary*. Auxiliaries are strictly grammatical words that do not hold semantic meaning. In Basque, verbal auxiliaries provide grammatical information about the tense, the mode and the person and number of the arguments of the action, the subject, the direct object and the indirect object.

B Prompts for the generative models

The prompts used for testing the generative models are shown in Table 11. They are a direct translation of the English prompts used in (Gao et al., 2021), which we show in Table 12 for completion purposes.

prompt	label
[premise], ezta? Bai, [hypothesis]	entailment
[premise], ezta? Ez, [hypothesis]	contradiction
[premise], ezta? Gainera, [hypothesis]	neutral

Table 11: Basque prompts used in the generative models.

prompt	label
[premise], right? Yes, [hypothesis]	entailment
[premise], right? No, [hypothesis]	contradiction
[premise], right? Also, [hypothesis]	neutral

Table 12: English prompts in English for XNLI.

C Native dataset guidelines for annotators

Translation to English: *The NLI (Natural Language Inference) task consists on classifying pairs of sentences according to their logical and semantical relation. The three possible relations are "Entailment" (when a sentence entails the other one), "Contradiction" (when both sentences contradict each other) and "Neutral" (when both sentences can either be true at the same time or not).*

We are trying to create a dataset for this task in Basque, and we need your help.

Your work consists on reading the sentences in the "Premise" column and writing three other sentences related to the first one. Only taking into account the first sentence and your own world knowledge, you should:

- *Write an entailment of the premise (a sentence that is true when the premise is true) in the "Entailment" column".*
- *Write a neutral statement in relation to the premise (a sentence whose truthfulness cannot be decided based on the premise) in the "Neutral" column.*
- *Write a contradiction of the premise (a sentence that is false when the premise is true) in the "Contradiction" column.*

If there is a problem with the premise, the row can be left blank, and the box in the "Problem" column must be checked.

We would like for the sentences to have some creativity, so we discourage the use of artifacts (for example, creating contradictions by simply adding "no" to the premise).

Example 1

Premise: The body language and the eyes were enough to communicate.

- *Entailment: Using body language and the eyes, they were able to communicate.*
- *Neutral: We human beings are able to communicate a lot of ways.*
- *Contradiction: To understand each other they had to talk.*

Example 2

Premise: Solte is one of those groups that sweat from minute one in their live performances.

- *Entailment: The group Solte are very lively in their concerts.*
- *Neutral: The group Solte gives a lot of concerts.*

- *Contradiction: Calmness is the main thing of the live performances of the Solte group.*

Original Basque: NLI (Natural Language Inference) ataza esaldi pareak sailkatzean datza, haien arteko erlazio lojiko eta semantikoan oinarrituta. Hiru erlazio aurreikusten dira esaldien artean: "Entailment" (esaldi batek bestea ondorioztatzen du), "Contradiction" (esaldiak kontraesankorrak dira) eta "Neutral" (esaldiek ez dute erlazio lojiko zuzenik).

Guk euskarazko NLI datu multzoa sortu nahi dugu, eta horretarako zure laguntza behar dugu.

Lan hau garatzeko "Premisa" zutabean dagoen esaldia irakurri behar da, eta esaldi horrekin erlazionatuta dauden beste hiru esaldi idatzi. Premisa esaldian bakarrik oinarrituz, eta zure munduko ezagutza kontuan izanik, gain zera egin behar duzu:

1. Idatzi premisaren ondorio bat (premisa egia denean egia den esaldi bat) "Entailment" zutabean.

2. "Neutroa" zutabean premisari buruzko esaldi neutro bat idatzi (premisa egia denean egia denik edo ez jakin ezin den esaldi bat).

3. “Contradiction" zutabean premisaren kontraesana idatzi (premisa egia denean faltsua den esaldi bat).

Erakutsitako premisarekin arazoren bat badago, lerroa hutsik utzi eta "problema" zutabeko laukian klik egin dezakezu.

Esaldi orijinalak nahi ditugu, sormena erakusten dutenak, beraz saiatu eskema berdinak ez erabiltzen (adibidez, kontraesanak sortzeko premisari "ez" hitza gehitzea).

Adibide 1

Premisa: Mimika eta begiak nahiko ziren komunikatzeko.

- Entailment: Mimika eta begiak erabiliz, komunikatzeko gai ziren.

- Neutral: Gizakiok hainbat komunikatzeko modu erabiltzeko gai gara.

- Contradiction: Elkar ulertzeko hitz egin behar zuten.

Adibide 2

Premisa: Zuzenekoetan izerdia lehen minututik botatzen duen talde horietakoa da Solte.

- Entailment: Solte taldekoak oso mugituak dira bere kontzertuetan.

- Neutral: Solte taldeak kontzertu asko ematen ditu.

- Contradiction: Lasaitasuna da nagusi Solte taldearen zuzenekoetan.