Basque and Spanish Counter Narrative Generation: Data Creation and Evaluation

Author: Jaione Bengoetxea Azurmendi

Advisors: Rodrigo Agerri

hap/lap

Hizkuntzaren Azterketa eta Prozesamendua
Language Analysis and Processing

Final Thesis

September 2022

Departments: Computer Systems and Languages, Computational Architectures and Technologies, Computational Science and Artificial Intelligence, Basque Language and Communication, Communications Engineer.

Acknowledgements

First of all, I would like to thank my supervisor, Rodrigo Agerri, for guiding and helping me throughout this year, for being patient with me as well as always pushing me to do better.
Additionally, I would like to express my gratitude to Yi-ling Chung for helping us understand the dataset and its division, as well as sharing their evaluation metric scripts with us.
I also want to thank my masters friends, because part of what made these years so full has been your presence in my life.
Finally, I want to thank my friends, Paula and Maria, for being a constant in this unpredictable times. Nothing would be the same without you.

Abstract

Counter Narratives (CN) are responses to Hate Speech (HS), which include non-negative feedback as well as fact-bound arguments, with the aim of de-escalating potentially hateful debates. However, due to the growing presence of the online world, HS quantity has been exponentially growing, and thus a need for automatic CN generation has been recently deemed necessary to deal with this hateful comments. Consequently, although research on this area has gained considerable interest in recent years, the majority of the studies have been focused on English. That is why the aim of this thesis is to provide some preliminary research on CN generation in Spanish and Basque, for which a HS-CN pair dataset will be used (CONAN). This dataset was Machine Translated (MT) both to Spanish and Basque, and each translated dataset was also manually post-edited. These datasets were used to conduct monolingual as well as crosslingual experiments, all of which were examined in terms of quantitative as well as qualitative evaluations. The results showed that, quantitatively speaking, the model trained with the Spanish post-edited datasets performed the best, while the MT model obtained the best results for Basque, although this outcome was highly influenced by training size. In terms of crosslingual results, the multilingual Basque model seems to slightly improve its monolingual baseline. Furthermore, the qualitative evaluation indicated that automatic metrics did not correlate well with human judgement, as manual evaluation showed a clear preference not only for the Spanish post-edited model, but also for the Basque post-edited experiment. This highlighted the importance of a manual evaluation step in text generation tasks.

Keywords: Counter Narrative generation, Spanish, Basque, crosslingual, post-edition

Contents

1 Introduction **1**

2 Related Work **4**
2.1 Hate datasets and hate detection 4
2.1.1 The definition of HS 4
2.1.2 Hate datasets 5
2.1.3 Hate detection 6
2.2 Countering of online hate 7
2.2.1 Effectiveness of CNs as an approach to fight Online Hate 7
2.2.2 Counter Narratives 8
2.2.2.1 Data collection methods 9
2.2.2.2 CN generation 10
2.3 Text generation in Spanish and Basque 11
2.4 Evaluation of NLG tasks 12

3 Methodology **14**
3.1 CONAN 14
3.2 Transformers 15
3.2.1 mT5 17
3.3 Evaluation metrics 18
3.3.1 BLEU 18
3.3.2 Rouge 19
3.3.3 Novelty 20
3.3.4 Repetition Rate 21

4 Basque Post-edition **23**
4.1 Post-edition statistics 23
4.2 Grammatical errors 24
4.3 Semantic related errors 26
4.4 Hashtags 26
4.5 Acronyms 27
4.6 Sentence level errors 28
4.7 Typos in original English dataset 29
4.8 Spanish post-edition 30

5 Experimental Setting **32**

6 Empirical Results **36**
6.1 Quantitative Evaluation 36
6.1.1 Monolingual experiments 36

6.1.2 Crosslingual results 42
6.2 Qualitative Evaluation 46
6.2.1 Monolingual results 47
6.2.2 Crosslingual results 48

7 Error Analysis **51**

8 Conclusions and Future Work **55**

List of Figures

1 Illustration of an attention layer 16
2 The Transformer architecture 16
3 Applications of T5 18
4 Dataset combinations for monolingual experiments (Spanish and Basque) . 33
5 Dataset combinations for crosslingual experiments 34
6 BLEU, Rouge-L and RR results with 5k training examples, except for eu-post 38
7 Novelty results with 5k training examples, except for eu-post 39
8 BLEU, Rouge-L and RR results with 2k training examples 40
9 Novelty results with 2k training examples 41
10 Crosslingual results for BLEU, Rouge-L and RR 44
11 Monolingual qualitative results 48
12 Crosslingual qualitative results and their monolingual baselines 49

List of Tables

1 Example of one CN-HS pair 14
2 Number of HS-CN pairs in CONAN 15
3 Basque post-edition statistics 23
4 Example of an incomplete sentence 29
5 Spanish post-edition statistics 30
6 Example of lack of gender concordance in Spanish 31
7 Train-test split 32
8 Monolingual results 37
9 Crosslingual results 43
10 Inter-annotator agreement (Cohen's kappa) 46
11 Qualitative results (average from annotators) 47
12 Summary of qualitative and quantitative results 51

1 Introduction

While the development of online social media has provided citizens with an unimaginable source of knowledge and communication opportunities, some people have found ways to use these new-found resources to spread hate. Hate Speech (HS), which has been a common human practice way before the rise of the online world, is defined by Davidson et al. (2017) as:

> "... language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group."

According to Silva et al. (2016), the targeted groups generally become victims due to their race, religion, behaviour, sexual orientation, physical appearance or gender.

This already present societal problem has exponentially escalated with the appearance of social media, as fighting HS online has become unmanageable due to its rapid increase in quantity. One of the main reasons behind this increase might be the anonymity social media provides, which facilitates the presence of HS without imminent consequences for the hate speaker (Reynolds et al., 2011).

In this context, researchers have pointed out the importance of having HS and its spread and impact under control: as stated in Benesch (2014), HS can be indicative of some potential preparation of an act of violence, and therefore early intervention could be extremely beneficial.

Consequently, some social media platforms have established some rules and regulations in order to control the spread of HS. The majority of social media companies currently have policies in their community guidelines condemning hateful content, for instance Twitter[1] or TikTok[2]. Not only that, but the European Commission, together with Facebook, Microsoft, Twitter and YouTube designed the *Code of Conduct on countering illegal hate speech online*[3] in 2016, to which several other companies later joined, such as Instagram or Snapchat.

It is not a coincidence that, around the same time, research on automatic HS detection gained a considerable interest, as this was deemed essential for the implementation of efficient HS policies in social media platforms. This is why several researchers provided HS datasets (Waseem and Hovy, 2016; De Gibert et al., 2018; Kolhatkar et al., 2020) as well as experimented with several configurations and model types for its automatic detection (Nobata et al., 2016; Davidson et al., 2017; Faris et al., 2020).

The progress on HS detection, together with progressively more strict social media policies has resulted in a fairly controlled online presence of HS. Nevertheless, the majority of the approaches social media platforms used in order to control online hate were based on deleting the HS or suspending or blocking the hate speakers, for which HS detection was incredibly useful. However, recent research has found that this might not be the best

[1]Twitter's code of conduct
[2]TikTok's code of conduct
[3]The EU Code of Conduct

approach (Benesch, 2014; Schieb and Preuss, 2016; Ernst et al., 2017), mainly because it goes against the principle of freedom of speech, but also because, according to Benesch (2014), the act of censoring the hate speaker seems to yield some disadvantages, such as:

- The state or the authority at hand could be biased, thus creating the risk of the laws or rules themselves being biased.
- There exists a possibility of backfire, as the audience of the hateful speech could increase or become even more extreme.

Given the potential complications approaches such as punishment could create, some other alternatives were considered in order to deal with hateful comments. By doing so, researchers found that Counter Narratives (CN) could serve as a great alternative to suspension or blocking of hate speakers, as Benesch (2014) indicated the effectiveness of CNs over censorship and deletion. Benesch et al. (2016) additionally create the most comprehensive taxonomy of CNs to date. The term CN refers to:

> A response to a hateful comment, which includes non-negative feedback as well as fact-bound arguments. (Chung et al., 2019; Benesch, 2014; Schieb and Preuss, 2016). In other words, a peaceful response to a violent comment, with the aim of de-escalating a potentially dangerous situation.

Over the years, research has proved the effectiveness of CNs as a response to HS, as opposed to other traditional approaches such as punishment (Ernst et al., 2017; Schieb and Preuss, 2016; Mathew et al., 2019). The main downside of this approach is that the creation of CNs is a rather time consuming and costly job, which the recent uprising of social media has further hindered. As a result, an automatic approach to CN generation has recently gained interest, as several studies have experimented on this field (Tekiroglu et al., 2020; Qian et al., 2019; Fanton et al., 2021).

However, the majority of the previous works on CNs have been centered around English, other than a few exceptions such as Chung et al. (2020), who worked on CN generation in Italian. Consequently, it comes as no surprise that no previous work has been done neither in Spanish nor in Basque on this area. That is the reason why the main goal of this thesis will be to explore the automatic generation of CNs both in Spanish and in Basque.

More specifically, this thesis will analyse the impact of the training corpus in CN generation in Spanish and Basque, comparing the generated output of Machine Translated (MT) data and post-edited training data. The research questions this thesis will focus on are the following:

RQ1 How does the training data size affect automatic metric results?

RQ2 In monolingual experiments, what are the differences in the generated output when models are trained with post-edited data, compared to Machine Translated (MT) data? Is it necessary to post-edit training data?

RQ3 In crosslingual environments, how do zero-shot models perform, compared to the monolingual baselines? Thus, how necessary is training data in the target language?

RQ4 How effective are data augmentation approaches such as monolingual MT datasets or crosslingual models, especially for low resource languages such as Basque?

RQ5 Following previous studies which raised concerns regarding the lack of correlation between automatic metrics and human judgement in Natural Language Generation (NLG), this thesis will assess the correlation between quantitative and qualitative evaluations.

Consequently, this thesis provides the following contributions:

- Provide the first HS-CN pair dataset in Spanish and Basque for CN generation, which serves both for monolingual or multilingual research purposes. These datasets are available at GitHub [4].
- Establish benchmark results for the automatic generation of CNs in Spanish and Basque.
- Introduce the first attempts at automatic text generation in Basque, to the best of our knowledge.
- Present a detailed comparison of generated CNs when trained with MT or post-edited datasets.
- Provide insights into crosslingual approaches (both zero-shot and multilingual) in Spanish and Basque.
- Propose important insights into data augmentation approaches and their effectiveness, which is especially relevant for low resource languages like Basque.

In order to do so, this thesis will follow the following structure: Section 2 will deal not only with CNs and its generation, but also HS and its detection, as well as commenting on text generation in Spanish and Basque and NLG evaluation approaches. Secondly, Section 3 will introduce the materials used in this thesis, such as the dataset or the evaluation metrics. Moreover, Section 4 will comment on the post-edition process of the Basque dataset, followed by Section 5, which will present the experimental setting. Both the quantitative and qualitative evaluations will be examined in Section 6, as well as an error analysis in Section 7. Finally, some conclusions and further research will be provided in Section 8.

[4] `https://github.com/ixa-ehu/conan-e`

2 Related Work

This section will introduce the state of the art on CNs and their automatic generation. In order to do so, the area of HS will first be explored, dealing with the difficulties behind its definition, the existing HS datasets and its detection. This is relevant, as CNs are responses to HS, and HS datasets and research on its detection is essential for CN generation. Following this, the effectiveness of CNs will be examined, as well as several CN data collection methods and previous literature on CN generation. This section will end with comments on NLG in Spanish and Basque, as well as the difficulty in evaluating these types of tasks.

2.1 Hate datasets and hate detection

2.1.1 The definition of HS

Although research on HS and its detection has gained a lot of interest in recent years, this has not always been the case. In fact, a common problem in the early years of this field was the lack of substantial HS datasets, which would then facilitate further research into, for instance, the automatic detection of HS. Consequently, interest considerably increased in the area of the collection and release of HS datasets.

However, one main difficulty was quickly identified, which was the lack of consensus when it comes to what exactly a HS involves (Ross et al., 2017; Waseem and Hovy, 2016). Preliminary research defined HS as a speech instance that abuses a person or a group of people due to a set of characteristics such as race, religion, sexuality or any other characteristic. Nevertheless, as research on this field increased, it was clear that more specifications were necessary to deal with the variations of HS definitions.

In other words, the range of HS definitions resulted in, for instance, some studies counting offensive language as HS, while others did not. One such work that directly addressed this was Davidson et al. (2017), who decided to establish the distinction between these concepts based on speaker intent: HS had to have the active intention of the speaker to humiliate or harm the targeted group, while offensive language used derogatory terms without the intention to do harm. This concept of speaker intention had been present since early on (Reynolds et al., 2011), but many researchers did not pay attention to it until Davidson et al. (2017). Many studies have since adopted this definition, and therefore this is how HS is understood in this thesis.

However, it is important to highlight that HS is a complex concept, greatly influenced by multidisciplinary factors such as society, making HS highly susceptible to change over time. Therefore, some concepts regarding HS are still up to debate, such as the number of speakers or targets involved in a hateful interaction. This is illustrated by both De Gibert et al. (2018) and ElSherief et al. (2018): while De Gibert et al. (2018) specified that a HS had to be directed to a group of people, ElSherief et al. (2018) decided to collect one-to-one instances of HS.

2.1.2 Hate datasets

Due to the fact that a generally accepted definition of HS is lacking, no common rules or processes have been established to create or collect hate datasets, creating a lot of disparity between studies. However, as interest recently increased in this field, a recent review of HS resources by Poletto et al. (2021) has highlighted the need for a common framework in order to avoid this issue in the future. Poletto et al. (2021) provides a very comprehensible collection of the most important HS datasets to date, as for example: (Waseem and Hovy, 2016; Ross et al., 2017; Hammer, 2017; Sprugnoli et al., 2018; Kolhatkar et al., 2020; Schäfer and Burtenshaw, 2019; Vidgen and Yasseri, 2020).

Most of these datasets used key-word based approaches to collect data, using words with negative connotations as well as collecting surrounding context of collected comments (ElSherief et al., 2018; Mathew et al., 2018). One of the most recent works on hate dateset collection has been done by Mathew et al. (2021), who presented a dataset which annotation included rationales, i.e. the spans in a post on which annotators based their labelling decision. By doing so, the bias and explanability of HS datasets could be explored.

Furthermore, Poletto et al. (2021) also provides an extensive list of HS datasets in languages other than English, which has been a very underdeveloped area until now. In fact, HS in other languages has considerably flourished in recent years, as several shared tasks started providing hate datasets in other languages, such as GermEval in German (Bai et al., 2018), EVALITA in Italian (Fersini et al., 2018a) or the more recent SemEval-2020 (Zampieri et al., 2020), which provided a multilingual datasets in Arabic, Danish, English, Greek, and Turkish. Poletto et al. (2021) also mentiones a considerable list of studies who collected hate datasets in various languages, as for example Sanguinetti et al. (2018); Ousidhoum et al. (2019).

However, as one of the languages this thesis focuses on is Spanish, it is of utmost importance to note that several shared tasks have provided Spanish hate datasets. For instance, IBEREVAL-2018 released a hate dataset in Spanish and English for the detection of misogynistic tweets, which allowed HS research in Spanish to begin, resulting in several multilingual studies such as (Goenaga et al., 2018; Pamungkas et al., 2018). An overview of this task can additionally be found in Fersini et al. (2018b). Furthermore, it needs to be highlighted that other shared tasks followed IBEREVAL-2018 in releasing more multilingual datasets in Spanish, such as the HatEval task in SemEval-2019 (Basile et al., 2019)

In fact, as these shared tasks started to solidify the ground for research in Spanish HS, some researchers started to collect and share more datasets. Apart from the shared tasks, two Spanish datasets have been provided in the last couple of years, to the best of our knowledge: Pereira-Kohatsu et al. (2019), who released a dataset consisting of 6000 expert-labeled tweets; and García-Díaz et al. (2021), who provided a balanced dataset of 7682 tweets for mysoginistic tweet detection in Spanish.

All in all, the recent appearance of a considerable number of hate datasets, not only in English but in various other languages, as well as the high number of submissions in several shared tasks just proves the growing interest of researchers in this field. Not only that, but

the fact that datasets in Spanish are starting to see the light shows the increasing interest in hate detection research in Spanish. A summary of common hate detection approaches will be discussed in the following section, together with an overview of the most recent research in Spanish.

2.1.3 Hate detection

Over the past decade, and especially in recent years, extensive work has been done regarding hate speech detection. Several works provide comprehensive reviews of existing studies such as Fortuna and Nunes (2018) or Schmidt and Wiegand (2017). Reviews such as Fortuna and Nunes (2018) stated that preliminary research in HS focused on classifiers based on features such as n-grams (Nobata et al., 2016), lexical resources (Gitari et al., 2015) or TF-IDF (Davidson et al., 2017), as well as algorithms such as Support Vector Machines (SVM) (Davidson et al., 2017) or Decision Trees (Burnap and Williams, 2016). In short, all these studies used traditional machine learning methods to build classifiers based on several features.

Nevertheless, Poletto et al. (2021), which is, to the best of our knowledge, the most recent review on HS detection, indicated that current studies have shifted their attention to deep learning approaches. Recent studies have worked on detection models based on architectures such as Convolutional Neural Networks (CNN) or Long-Sjprt Term Memory (LSTMs), such as Badjatiya et al. (2017), as well as double deep learning approaches such as Faris et al. (2020). Not only that, but transfer learning methods have gained considerable interest in recent years, as pre-trained language models such as BERT have been released. These approach has been found to perform exceptionally well in current research (Dowlagar and Mamidi, 2021).

As we can see, more and more studies are focusing on the task of HS detection, not only in English but in other languages as well (Akhter et al., 2020; Al-Hassan and Al-Dossari, 2021). However, when it comes to HS detection in Spanish, it is undeniable that it is still in its infancy, as rather few works have dealt with this topic.

Following traditional machine learning approaches, for instance, García-Díaz et al. (2021) worked on misogyny detection and found their best results to be from a type of SVM model, which was based on Average Word Embeddings and Linguistic features. One of their future works pointed out the exploration of deep learning methods. In fact, there are several studies who utilize deep learning methods, such as (Plaza-Del-Arco et al., 2020), who trained LSTM models and argued that the performance was low due to the low amount of training data; or Pereira-Kohatsu et al. (2019), who used a double deep learning approach, combining LSTMs and Multilayer Perceptron (MLP), as well as using token embeddings, emojis and word expressions as input, enriched by TF-IDF.

Furthermore, following research in other languages, the most recent works on Spanish HS detection have presented novel work using pre-trained language models such as BERT or XML. In fact, Plaza-del Arco et al. (2021) compared the results obtained with the multilingual models mBERT and XML to the Spanish monolingual BETO, finding that BETO outperformed both multilingual language models, thus highlighting the importance

of pre-trained language models in Spanish. The most recent attempt at Spanish HS detection, to the best of our knowledge, is by (García-Díaz et al., 2022), who concluded that a combination of linguistic features and transformer architecture models (such as BETO) resulted in improved state of the art performance.

Having explored the literature behind HS, the next section will present a specific type of response for hateful comments: counter speech.

2.2 Countering of online hate

2.2.1 Effectiveness of CNs as an approach to fight Online Hate

In recent years, authorities and researchers have been looking for alternatives for blocking or deleting HS content, as this approach goes against the right of freedom of expression. It is in this context where responding to hateful comments with counter speech was proposed. As previously mentioned, counter speech, also referred to as CN, is defined as a non-violent, fact-based argument, which is presented with the aim to de-escalate a potentially violent situation or confrontation (Benesch, 2014; Schieb and Preuss, 2016). Consequently, we could say that while punishment such as blocking or deleting HS comments focuses on the speaker and the audience of said speech, countering focuses on the speech act itself (Benesch, 2014).

As this is a rather new concept, some research has been done on the good practices for acceptable CN creation (Benesch et al., 2016). In fact, these studies were extremely helpful to introduce this novel concept, as some guidance on what to do or not to do was offered: while strategies like warning of consequences or showing empathy and affiliation had positive results in CN effectiveness, other methods such as hostile or aggressive tones and harassment and silencing did not, and were thus advised to be avoided.

This preliminary research on CNs already showed positive results in terms of effectiveness, showing that CNs could be a more effective and improved alternative to fight HS. This opened the door for more studies on the field: works mainly studied counter speech and its effect on mitigating online hate, by providing extensive analysis of CNs as responses to HS, and showed generally good results in favour of counter speech. Different social media platforms have been studied in a variety of different works: Facebook (Schieb and Preuss, 2016), Youtube (Ernst et al., 2017; Mathew et al., 2019) or Twitter (Wright et al., 2017; Mathew et al., 2018).

Schieb and Preuss (2016) presented their findings on the effectiveness of counter speech on Facebook, through a simulation. They found that counter speech had in fact been able to have an effect on hate speakers, but its level of success strongly depended on the number of hate speakers it reached to: generally, the influence of the counter speech increased as the number of hate speakers decreased.

Regarding Youtube, Ernst et al. (2017) analysed comments in counter speech videos on the topic of Muslims and Islam, and indicated that there was a potential motivation in people to discuss topics presented in the media, which could lead to "increasing knowledge of the presented facts". Mathew et al. (2019) released the first counter speech dataset

by extracting direct comments from Youtube videos, annotated into counter/non-counter comments. They analysed the collected CN instances, highlighting that different communities use different counter speech strategies, as well as indicating that counter speakers and non-counter speakers differ in language.

Regarding Twitter, Wright et al. (2017) conducted a qualitative research on the influence of counter speech on Twitter, and found that counter speech could influence a change in attitude in the hate speaker, as for instance an apology. Another study on Twitter was conducted by Mathew et al. (2018), as they provided a dataset of tweet/reply pairs scraped from Twitter, where the tweet was a HS and the reply a counter speech. Their analysis of the collected data brought some interesting findings, such as hate tweets written by verified accounts being more likely to go viral, and HS and CN accounts being distinguishable by the language that they used in their tweets (similar to Mathew et al. (2019) on Youtube).

In short, these studies have proved the effectiveness of CNs as a response to HS. That is why CNs have gained a considerable interest, and research has started investigating the collection and automatic generation of counter speech, which is explored in the following section.

2.2.2 Counter Narratives

Preliminary work on CNs included the automatic detection of said speech: Mathew et al. (2019) worked on the first empirical work on CNs and attempted its automatic detection by building some general classifiers such as SVMs as well as neural models such as LSTMs, following HS detection studies. Not only that, but Mathew et al. (2018) also built a classifier to distinguish HS from CNs, with an accuracy of 78%.

Nevertheless, as research on this field gained considerable interest, some researchers pointed out the importance of investigating the area of CN generation instead. This could bring a lot of benefits to the very complicated task of HS fighting: with the rise of online social media, the amount of HS has become unfeasible, and thus automatically generating CNs would greatly aid this problem.

In this context, however, the first obstacle in this research area was quickly identified: in line with studies on HS, there was a general lack of good quality, publicly available CN datasets. The very few available datasets where either incomplete, as they did not include the HS instances (e.g. Mathew et al. (2019)), or the data was collected through crawling methods (e.g. Mathew et al. (2018)). Crawling methods involve collecting data through social media platforms, and posteriorly conducting a manual annotation process. This method has been pointed out to perform rather poorly regarding data quality for the specific task of CNs, as for instance by Chung et al. (2019), who indicated several disadvantages:

- Datasets are often ephemeral, as online content is fleeting and content deletion results in the vast majority of the datasets being lost.
- The data collection process often involves some kind of template for data detection (e.g. Mathew et al. (2018)). This limits the richness of both HS and CNs.

- Some responses contain hostile language, which goes against the definition and intent of CNs, and negatively impacts their effectiveness.

In short, this scarcity of good quality datasets has been an obstacle for the growth of research in the area of CN generation. However, some recent studies have proposed various collection methods that will be presented below.

2.2.2.1 Data collection methods

In this context, in order to overcome the limitations that crawling collection methods had, several studies proposed different data collection methods, such as through crowdsourcing (Qian et al., 2019), nichesourcing (Chung et al., 2019) or hybrid methods (Tekiroglu et al., 2020; Fanton et al., 2021), with the aim of obtaining appropriate datasets for CN generation tasks.

Crowdsourcing In this type of collection method a computationally difficult task is conducted by a group of people, usually gathered through social networks, or as Boer et al. (2012) stated, by the "faceless" crowd. One work which used this method of data collection was that of Qian et al. (2019). Their dataset consists of online HS instances collected through a keyword search, followed by a crowdsource-led manual creation of CNs for each of the previously collected HS.

Nichesourcing This approach involves complex computational tasks which are completed by experts in the specific field of the required task (Boer et al., 2012). This is how Chung et al. (2019) created CONAN (COunter NArratives through Nichesourcing), a non-ephemeral, expert-based and multilingual HS/CN pair dataset. Given the nature of CNs, as it is a rather specialised term, data collectors often need training to fully understand what exactly they involve. Therefore, using experts on the field such as non-governmental organization (NGO) operators as in Chung et al. (2019) seems a reasonable choice. However, data scarcity was one of its shortcomings, for which CONAN underwent a data augmentation process.

Hybrid methods In the midst of the scarcity of HS/CN data for NLG tasks, some researchers have been exploring hybrid data collection methods, where part of the data is manually collected and part is automatically generated. One such work is that of Tekiroglu et al. (2020) as they proposed a novel method to collect CNs by automatically generating them through an author-reviewer architecture. In short, they found that the model trained with nichesourcing data obtained the most diverse results, both in terms of quantitative metrics, but also regarding diversity and novelty.

Another study has been recently conducted on a similar line: Fanton et al. (2021) worked on CN collection through generation, basing their work in the author-reviewer architecture of Tekiroglu et al. (2020), but with a difference: they conducted several iterations of the data collection process, and in each interation, the generated CNs post-edited

by the reviewers would be added to the training data of the language model in the next iteration. Their findings suggest that data collection through iteration favours efficiency (in terms of acceptance rate of generated CN and post-editing effort), but hinders data quality (regarding diversity and novelty).

2.2.2.2 CN generation

In this context, and with the benefit of these new data collection methods, CN generation tasks have gained a lot of interest. Consequently, several studies have been working on various different aspects of CN generation: (i) various CN generation model types have been explored (Qian et al., 2019; Pranesh et al., 2021) (ii) some other works have dealt with the lack of diversity and relevance of generated CNs (Zhu and Bhat, 2021; Chung et al., 2021) (iii) some preliminary research on languages other than English has arisen (Chung et al., 2020).

CN generation models Qian et al. (2019), apart from providing a crowdsourced-dataset, additionally worked on the automatic generation of counter narratives, as they trained three encoder-decoder based generation models. They provided a benchmark for future generation studies, as their results indicated a clear room for improvement. Pranesh et al. (2021), who used the same dataset as Qian et al. (2019), also worked on CN generation by fine tuning three pre-trained language models. Their results showed that while one of the models performed the best in the quantitative metrics (DialoGPT), another one of their models was the one obtaining the best results in the manual evaluation (BART). This goes in line with Qian et al. (2019), who found that although quantitative results showed good results, human evaluation showed a clear preference for human created CNs. This indicates a lack of human correlation in automatic metrics, which will be discussed in Section 6.2.

More recently, Tekiroglu et al. (2022) also focused their CN generative research on pre-trained language models, and found that autoregressive models such as GPT-2 and DialoGPT performed the best in terms of specificity and novelty. They additionally proposed a pipeline where an automatic post-edition step would be added after the generation of CNs, in order to refine the generated output.

Lack of diversity and relevance Most CN generation works have a general shortcoming, which is the lack of diversity and relevance in their generated text (Fanton et al., 2021; Qian et al., 2019). This is why several studies have focused on these more task specific aspects of CN generation, such as Zhu and Bhat (2021), who worked on CN generation using both Qian et al. (2019)'s dataset, as well as part of the English section of Chung et al. (2019)'s corpus. They propose a three-module pipeline which they called "Generate, Prune, Select" (GPS). This involved the generation of phrases, followed by the pruning of the ungrammatical sentences, and finishing with the selection of the most relevant responses to the given HS. Their findings show that GPS performs better both in terms of diversity and relevance when compared to several baseline models.

Furthermore, Chung et al. (2021) introduced a knowledge-bound CN generation pipeline, which included a key-phrase generation step before actual CN generation through pre-trained language models. They found that having this extra information provided by key-phrases in the CN generation step resulted in suitable and informative CNs.

CN generation in other languages The fact that all this previous work has been done solely in English needs to be highlighted. One work dealing with CN generation in other languages is Chung et al. (2020), who decided to work on CN generation in Italian by studying the effects of using silver data compared to golden data, as well as a combination of both. By doing so, Chung et al. (2020) found that using silver data as a form of data augmentation, combined with golden data and a strong language model like GePpeTto resulted in promising outcomes. This is specially interesting for low resource languages, where the collecting of gold data is often costly and difficult.

As we can see, research on CNs in languages other than English is extremely scarce. To the best of our knowledge, no previous work has been done in Spanish or Basque regarding CNs or their generation, which are the languages this thesis will work on. Due to the lack of works in this area, the next section will explore text generation studies in Spanish and Basque, as these works are deemed to be a good starting point for CN generation research.

2.3 Text generation in Spanish and Basque

As previously mentioned, studies have started to arose regarding HS in Spanish. However, as this research line in Spanish is rather novel, the variety of studies and topics covered is quite limited, and thus CN generation in Spanish is yet non-existent.

Nevertheless, taking into account that CN generation is a NLG task, the research field that most assimilates to this type of task in Spanish is that of abstrative summarisation. Although being an underdeveloped line of research, some attempts have been made in the last couple of years to automatically generate summaries: the majority of these attempts have been made in multilingual works, such as Cao et al. (2020); Hasan et al. (2021); Varab and Schluter (2021), where multilingual pre-trained language models have been extensively used. All in all, these studies have highlighted the importance of transfer learning through multilingual language models such as mBART or mT5, especially for low resource languages, as transfer learning has been proved to aid data scarcity.

However, although these studies deal with a variety of languages, Spanish included, they do not focus on the language and domain specificities of Spanish. The first study on abstractive summarisation that focused on Spanish, to the best of our knowledge, is by Esteban and Lloret (2017), who created a web interface that would generate abstractive summaries of reviews in the travel industry.

Furthermore, the first monolingual attempts in Spanish was conducted by Ahuir et al. (2021): Their aim was to compare the performance of monolingual and multilingual models, and thus trained monolingual models in Spanish (NASES) and Catalan (NASCA), which were based on the BART architecture. They obtained promising results, as the monolingual models for both Spanish and Catalan performed slightly better than their multilingual

baselines (mBART), obtaining higher level of abstractiveness in their summaries. They especially highlighted the importance of monolingual models for low-resource languages like Catalan, as they are usually under-represented in the training corpora of multilingual language models, thus decreasing their performance.

Finally, it is important to note that this thesis not only deals with Spanish CN generation, but it also explores this same task in another language: Basque. As it can be seen by the state of the art of this field, no previous research has been done on Basque HS detection, and thus no studies have dealt with CN generation either. Not only that, but to the best of our knowledge there is also no previous literature in NLG in Basque, as not even the previously mentioned multilingual abstractive summarisation works have included Basque.

2.4 Evaluation of NLG tasks

As previously mentioned, this thesis deals with a NLG task, more specifically, CN generation. However, when it comes to evaluating text generation tasks, little to no agreement can be found regarding common practices or effectiveness of the used evaluation methods, as it is pointed out in several review works on this topic such as Celikyilmaz et al. (2021); Sai et al. (2020). Until recently, human evaluation has been deemed to be the common practice, and used as the gold standard for new automatic metric developments. However, these manual methods are costly and time consuming, which is why the need to develop automatic metrics that provide human-approved quality is of utmost importance.

The problem behind the use of automatic metrics in NLG tasks is that this is a rather complex field, where several facts need to be taken into consideration: firstly, text generation evaluations should have a certain level of explainability, which is extremely important to identify possible errors and learn how to correct them. Determining where the errors stem from could greatly help the improvement of text generation systems, as well as increase users' trust in the generated output, as stated by Celikyilmaz et al. (2021).

The second concept that needs to be taken into consideration in text generation evaluation is that NLG includes several different types of tasks (e.g. question answering, text summarisation, data-to-text generation, etc.), and each task has its own nuances and specific characteristics that need to be checked. Nevertheless, there are no metrics for every specific NLG task, and thus some studies used metrics not tailored for their specific generation tasks. For example, BLEU, an n-gram based metric initially provided for MT, has recently been widely used for a variety of NLG tasks, but has been found not to correlate very well with human judgement in tasks such as dialogue response generation.

In fact, the lack of human correlation with the automatic metric results is one of the most common difficulties of NLG evaluation, as it can be observed in the abovementioned reviews (Celikyilmaz et al., 2021; Sai et al., 2020). One way to overcome this would be to create task specific metrics, but as mentioned in the previous paragraph, this would not be feasible. Celikyilmaz et al. (2021) mention an alternative, which is the creation of improved corpus quality. They state this could help improve human correlation, as well as decrease bias in the generated output. The lack of human correlation has been

highlighted by several researchers in the CN generation tasks, as for instance Qian et al. (2019); Pranesh et al. (2021).

In short, NLG automatic metrics still suffer from a great deal of variation and shortcomings, which is why analysing text generation performance both with automatic metrics and human evaluation is advised. By doing so, it is possible to detect whether the automatic metrics correctly reflect human judgement results or not.

Summary This section has introduced previous literature both in terms of HS and its detection, as well as CN effectiveness and its generation. The problem of the majority of the previous research being in English has been highlighted, as no research has been found on CNs in Spanish and Basque. As this thesis focuses on CN generation in these two languages, previous work on NLG tasks in Spanish and Basque has been presented, indicating that text generation in Spanish is rather under-researched, and non-existent in Basque. Finally, the problem of correlation of automatic metrics with human judgement in NLG tasks has been described.

3 Methodology

This section will introduce the materials and methods used in this thesis, by first presenting the dataset used to build the models, followed by the description of the transformer architecture. This section will end with the examination of the automatic evaluation metrics employed in the experiments in Section 6.1.

3.1 CONAN

CONAN (COunter NArratives through Nichesourcing) is a dataset containing a series of HS instances, together with their corresponding CNs, on the topic of Muslims and Islam. The motivation behind CONAN's creation was, according to Chung et al. (2019), to provide a non-ephemeral, expert-based, multilingual corpus of HS-CN pairs.

In order to do so, Chung et al. (2019) decided to step aside from traditionally used data collection methods (such as crawling (Mathew et al., 2018) or crowdsourcing (Qian et al., 2019)) and used Boer et al. (2012)'s proposal of nichesourcing, where complex computational tasks are completed by experts in the specific field of the task. In this case, CONAN was created through data collection sessions with NGO operators, i.e experts on countering hate online. Through this collection approach, they were able to create a non-ephemeral, expert-based corpus, not only in English, but also in French and Italian.

The main reason behind this choice (nichesourcing over crowdsourcing) was the increasing importance of quality of the output: as stated by Tekiroglu et al. (2020), data collection through crowdsourcing provides enough data for deep learning approaches, but the data is simple and stereotypical. However, even though nichesourcing generates less data, it appears to be more diverse and less stereotypical, as it is generated by experts in the field, thus suggesting better end results in terms of quality. An example of a HS-CN pair in CONAN can be observed in Table 1:

Hate Speech	**Counter Narrative**
Islamic are criminals: they rape, enslave and murder people. Islam is more a worship than a religion and we do not have anything to share with them.	The myth that Muslims are dangerous and violent is a product of our vilifying media. Don't believe everything you read.

Table 1: Example of one CN-HS pair

However, nichesourcing has a distinct disadvantage: collecting data through experts is time consuming and costly, and the final data quantity is considerably lower than that of other collection methods. Therefore, in order to overcome this quantity problem, Chung et al. (2019) carried out a data augmentation process. After the initial data collection sessions, they obtained several CNs for the same HS, so they had several non-expert workers create two paraphrases of each original HS, which were then paired with CNs of the original

HS. Through this process, the unbalance between HS-CN was fixed, as well as augmenting the overall corpus. Evidently, the quality of the augmented pairs was manually tested, concluding that these pairs were almost as good as the original pairings.

Furthermore, the original pairs in French and Italian were translated into English, in order to have a parallel corpus of language pairs. By doing so, cross-language approaches are facilitated, while additionally augmenting the number of HS-CN pairs in English.

	English	**French**	**Italian**	**Total**
Original	1288	1719	1071	4079
Augmented	2576	3438	2142	8159
Translated	2790	-	-	2790
Total	6654	5157	3257	15025

Table 2: Number of HS-CN pairs in CONAN

Regarding the quantitative description of the corpus, the number of HS-CN pairs can be observed in Table 2. The initial number of HS-CN pairs in all three languages was of approximately 4000, but after the data augmentation process and the translation of French and Italian data to English, the final total number of pairs was increased to around 15 thousand. In terms of total number of pairs per language, English, Italian and French pairs were initially 1288, 1719 and 1071 respectively, while by the end, these numbers were increased to 6654, 5157 and 3257. This is a significant increase and demonstrates how beneficial the augmentation and translation processes were.

3.2 Transformers

Transformers are a novel architecture first introduced by Vaswani et al. (2017) in order to deal with sequence-to-sequence (seq2seq) tasks. Before transformers, seq2seq tasks were handled by deep learning methods such as Recurrent Neural Networks (RNN), which could also include an attention mechanism. This is a mechanism that allows the decoder to go back and pay special attention to particular parts of the input sequence, which was introduced by Bahdanau et al. (2014), specifically for MT.

This concept was introduced through an example which can be observed in Figure 1, where a sentence is translated from English to French. In this case, word-by-word translation would not work, as sentence order in the two languages is different. Attention aids this situation, as the output word 'européenne' was predicted by paying attention to both input words 'European' and 'Economic', being able to capture the word order difference from English to French.

Despite the introduction of this attention mechanism, RNNs still had some disadvantages: firstly, RNNs were rather hard to train, as the input was introduced sequentially (word by word) (Goled, 2021). Secondly, and more importantly, these models still had problems dealing with long-range dependencies (Goled, 2021), as they have a limited window which can be stored in memory. Therefore, the longer the sentence, the less the model

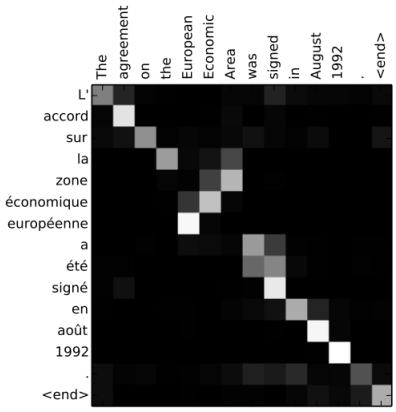

Figure 1: Illustration of an attention layer

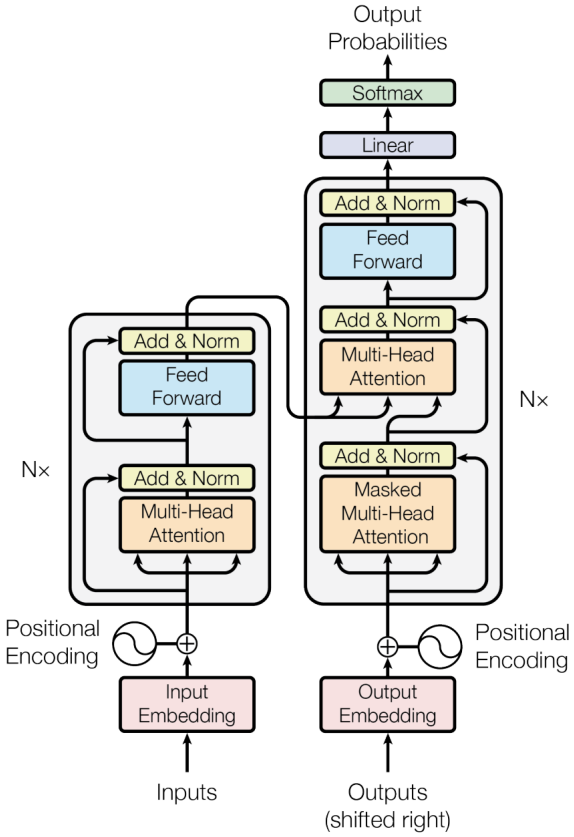

Figure 1: The Transformer - model architecture.

Figure 2: The Transformer architecture

will be able to remember from the beginning of the sentence.

In this context, Transformers were introduced to overcome these problems: to begin with, transformers can process input parallely, thus being able to process the whole sentence at once rather than sequentially, like RNNs. Furthermore, these architectures also introduced the so-called self-attention layers in order to deal with the long-range dependency problem (Goled, 2021). Self-attention is defined by Vaswani et al. (2017) the following way:

> "Self-attention, sometimes called intra-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence."

In other words, self-attention assists the model in understanding the input word in the context of the surrounding words, providing the model with a higher level of internal representation of language (Markowitz, 2021).

The entirety of the transformer architecture is illustrated in Figure 2, which was provided by Vaswani et al. (2017), where we can see how both the encoder on the left and decoder on the right, as well as their attention layers which help solve the memory problem previously mentioned. This thesis made use of a transformer architecture called mT5 to conduct the experiments, which is introduced below.

3.2.1 mT5

mT5 is a multilingual variant of T5 (**T**ext-**t**o-**T**ext **T**ransfer **T**ransformer), in short defined as a pre-trained language model that uses a basic encoder-decoder Transformer architecture, which deals with NLP text-to-text tasks, as implied by the name. T5 was trained on web extracted text from Common Crawl, which was first preprocessed to obtain the cleanest data possible, as well as to discard texts in languages other than English. The final cleaned text-corpora was named the **C**olossal **C**lean **C**rawled **C**orpus (C4) by Raffel et al. (2019), which was the final dataset used to train T5.

This multilingual version of T5 was introduced by Xue et al. (2020), which was heavily based on T5, as it maintained the majority of its characteristics, such as its text-to-text format. mT5 was trained with a multilingual variant of C4, which contained text in 101 languages, also obtained from the Common Crawl web. It followed T5's text-cleaning criteria to obtain a cleaned mC4 dataset, which was used to train mT5.

Some of the main applications of T5 (and therefore of mT5) can be observed in Figure 3, which was established by Raffel et al. (2019). As we can see, T5 can be utilised in several text-to-text tasks such as machine translation, text classification, regression tasks such as text similarity, or document summarisation. This makes mT5 a rather good candidate for the task of this thesis: CN generation, which works as a response to an input text, namely a HS. In short, a text-to-text task where the input is a HS and the output is an automatically generated CN which works as a direct reply of said input.

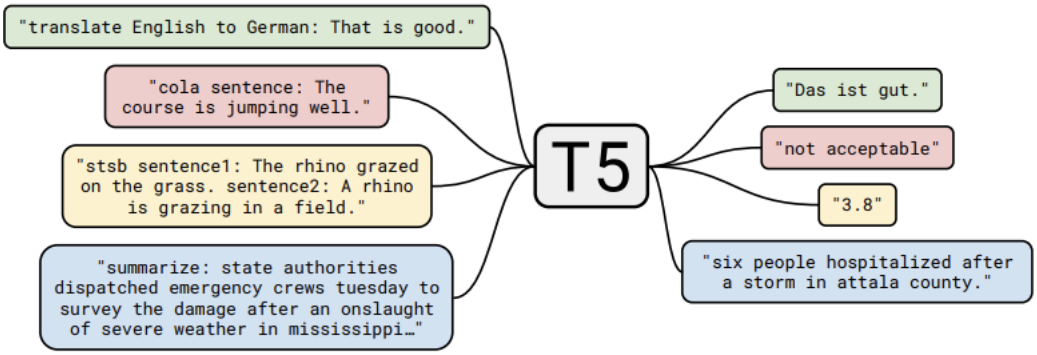

Figure 3: Applications of T5

3.3 Evaluation metrics

The following section introduces the four automatic metrics that will be used to quantitatively evaluate the experiment results: BLEU, Rouge, Novelty and Repetition Rate (RR).

3.3.1 BLEU

BLEU (Bilingual Evaluation Understudy) is a precision based evaluation metric originally developed for MT evaluation by Papineni et al. (2002). In other words, this metric evaluates a MT sentence by looking at the n-gram correlation between the translated sentence and the reference.

$$preicision = \frac{number\_of\_overlapping\_words}{total\_words\_in\_candidate\_translation} \tag{1}$$

However, solely using precision to evaluate MT has a clear downside, which can be illustrated with the following example:

Spanish: todos los musulmanes son terroristas

Reference Translation: all Muslims are terrorists

MT Candidate: all all all all all

As we can see here, the quality of the machine translated sentence is evidently low. However, if we were to calculate its precision, we would realise that the result is actually 1, providing us with the false impression that the MT sentence is in fact perfect:

$$precision = \frac{1+1+1+1+1}{5} = 1 \tag{2}$$

BLEU deals with this matter by calculating a modified n-gram precision instead of the traditional precision. In order to do so, the first 'all' in the machine translated sentence matches with the first 'all' in the reference sentence. Once a word in the reference sentence

matches with a word in the MT sentence, it is excluded, and thus cannot be matched again. By using this technique, the precision will be computed the following way:

$$modified\_unigram\_precision = \frac{1+0+0+0+0}{5} = 0,2 \tag{3}$$

This result correlates better with human judgment, and it is what BLEU is based on. This can evidently be recreated with any other n-gram, and according to Papineni et al. (2002), when a translated sentences has a high number of word matches with the reference (unigram), it also means that the adequacy is high. Not only that, but when longer n-gram matches result in a high precision, it usually also means a high fluency.

However, the problem with modified n-gram precision is that it tends to favor short sentences, which is why BLEU implements a brevity penalty:

$$PB = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } c \leq r \end{cases} \tag{4}$$

By doing so, "a high-scoring candidate translation must now match the reference translations in length, in word choice and in word order" (Papineni et al., 2002). In other words, a candidate translation sentence which is too short will now be penalized.

3.3.2 Rouge

Rouge (Recall-Oriented Understudy for Gisting Evaluation) is a recall based evaluation metric which was originally designed for summary evaluation by Lin (2004). When calculating Rouge, both the precision and the recall, as well as the F1-score are computed. In other words, by looking at the recall we are observing how much of the reference summary there is in the candidate summary. Not only that, but the precision tells us how much of the candidate summary is actually relevant, i.e. how much of it also appears in the reference summary. By doing so, we ensure that no extra information is being added, and we avoid excessively long summaries.

$$recall = \frac{number\_of\_overlapping\_words}{total\_words\_in\_reference\_summary} \tag{5}$$

$$precision = \frac{number\_of\_overlapping\_words}{total\_words\_in\_candidate\_summary} \tag{6}$$

There are several different types of Rouge, two of which are presented below: Rouge-N and Rouge-L.

Rouge-N It calculates the number of matching n-grams in the reference and candidate summary. For instance, we can compute Rouge-1 by counting the matching unigrams from the following example:

Reference: all Muslims are violent people

Candidate: all Muslims use violence to hurt people, are bad

If we were to calculate the unigram recall and precision of these sentences, we would observe the following:

$$Rouge1\_recall = \frac{4}{5} = 0,8 \tag{7}$$

$$Rouge1\_precision = \frac{4}{9} = 0,44 \tag{8}$$

By looking at these results, we can see that 4 words out of 5 from the reference were present in the candidate summary, which, according to recall, means that almost all the information from the reference is present in the candidate. However, the precision being that low means that many of the words in the candidate summary were not present in the reference, i.e. they are not relevant. This is a very good example of why both precision and recall are needed.

Rouge-L It finds the Longest Common Subsequence (LCS) between the reference and the candidate, and computes the precision and recall based on the LCS. For instance, with the following example:

Reference: Muslims, go back home.

Candidate: I do not think Muslims belong here, I say we send them back home

The LCS in this example consists of three words, 'Muslims back home'. Therefore, precision and recall for Rouge-L in this example would be computed as follows:

$$RougeL\_recall = \frac{3}{4} = 0,75 \tag{9}$$

$$RougeL\_precision = \frac{3}{14} = 0,21 \tag{10}$$

One of the advantages of taking the LCS into account when calculating precision and recall is that, unlike Rouge-N, Rouge-L is not dependant on consecutive n-gram matches, but instead captures in-sequence matches. This means that Rouge-L captures the overall sentence structure and word order more accurately than other types of Rouge.

3.3.3 Novelty

Novelty is a quality metric based on the Jaccard similarity function (Wang and Wan, 2018). While the Jaccard similarity function in itself measures how similar two sets are, novelty is rather the opposite: as stated by Tekiroglu et al. (2020), "novelty is used to measure how different the generated sequences are with regard to the training corpus". Thus, we could say that novelty is, in short, the normalized jaccard similarity score.

In order to compute this, first of all, the vocabulary from the input texts needs to be retrieved by removing any punctuation from the training and generated CNs. After obtaining the vocabulary from both input and output texts, the normalized Jaccard similarity

can be calculated: we need to obtain the unigram-based intersection from both texts, as well as the unigram-based union, and then divide the former by the latter. Finally, the Jaccard similarity score is normalized (1 - obtained_score), consequently calculating our novelty metric. In other words, by normalizing the Jaccard similarity score, our novelty metric tells us how different the two CNs are, whether the generated CNs are novel or not. This is illustrated in 12.

$$jaccard\_similarity = \frac{unigram\_based\_intersection}{unigram\_based\_union} \tag{11}$$

$$Novelty = 1 - \Big(\frac{unigram\_based\_intersection}{unigram\_based\_union}\Big) \tag{12}$$

Therefore, taking into account the following example:

> Training CN: Islamic schools in Birmingham are still separating students based on gender, which is illegal
>
> Generated CN: gender segregation in faith schools is clearly a negative, but hardly exclusive to some Islamic schools

The novelty score would be computed the following way:

$$Novelty = 1 - \Big(\frac{5}{24}\Big) = 0,79 \tag{13}$$

In the example above, we can observe only five words both in the training and predicted CN: "Islamic", "schools", "are", "gender" and "is". After dividing this intersection with the union of unigrams from both texts and normalizing it, we get a novelty score of 0,79. Obtaining such a high novelty result as this one means that the dissimilarity between these texts is quite high, thus the generated text is rather novel.

Consequently, if the novelty score is low, the generated CNs are closer to being copies of the original text rather than a generation. On the contrary, if the novelty score is high, then the texts are indeed different, i.e. novel, as in the example above. It needs to he highlighted that this could mean creativeness, but it could also mean lack of relatedness. Therefore, this metrics, as many others, needs to be contextualised with respect to a user-based or human evaluation.

3.3.4 Repetition Rate

Repetition Rate (RR) is a metric that calculates the level of repetitiveness in texts, introduced by Bertoldi et al. (2013); Cettolo et al. (2014). It is computed by calculating the rate of non-singleton n-grams that are repeated in a text. Furthermore, it is calculated on fixed-size sliding windows, thus allowing the comparison between corpora of different sizes. Not only that, but as stated by Cettolo et al. (2014), another advantage of sliding windows is that the original sequence of the text is preserved and thus its linguistic features.

Therefore, this thesis would look for the lowest possible RR, as this would mean that the generated CNs are not simply repeated sentences. In other words, a high RR score would mean that there is a large number of repeated CNs, which would negatively impact the quality of the generated CNs.

Summary This section has introduced the materials used in this thesis in order to conduct the experiments described in Section 5. Firstly, the dataset by Chung et al. (2019) has been presented, which is what we have used to build out models. Secondly, the transformer architectures have been explores, as well as the pre-trained languge model used to fine tune our experiments: mT5. Finally, the automatic evaluation metrics used to measure the quantitative quality have been defined.

4 Basque Post-edition

This following section will examine the post-editing process that was carried out, in order to obtain the Spanish and Basque post-edited version of the CONAN dataset. This dataset, originally containing English, French and Italian data, was first MT to Spanish and Basque, using the Google API[5], for both the train and test set. These MT datasets were stored to perform experiments on them later.

Besides automatically translating the corpus, both the Spanish and Basque datasets underwent an additional process of manual post-editing. The Spanish MT dataset was post-edited in its entirety by a student intern. The Basque dataset, however, underwent two separate post-editing occasions: firstly, the training dataset was post-edited in the span of 3 months, although only 2000 out of totality of HS-CN examples in the train set were post-edited. Secondly, the test set was also post-edited by professional translators at Elhuyar, but this only included CNs, leaving the HS in the test set unedited.

Generally speaking, during the post-edition process of the Basque training set, three types of post-editing instances could be identified: (i) instances where no edits where made, as the MT sentence was correct (ii) minimal changes were needed, where one word, termination or punctuation had to be changed. For instance, the noun 'islam' was often translated as 'islam', but in the majority of the contexts in Basque this word needs the post-position '-a': 'islama'. And finally, (iii) instances where more substantial changes were made, i.e whole expressions, word order, or the addition/deletion of a whole sentence.

This following section will comment on the Basque post-edition process and peculiarities found, informing about quantitative as well as qualitative aspects of the post-edition process.

4.1 Post-edition statistics

This section describes the quantitative statistics found in the Basque corpus, commenting on the number and percentage of post-edited HS and CNs in the train set, as well as post-edition statistics for the CNs in the test set.

	Unique	**Post-edited**	**Percentage (%)**
Train-HS	205	148	72.20
Train-CN	986	759	76.98
Test-HS	136	-	-
Test-CN	1268	926	73.03

Table 3: Basque post-edition statistics

As we can see in Table 3, the percentage of post-edition both in the train set and test set is rather high, as all of them are above 70%. In other words, these statistics seem to imply that the majority of the Basque MT instances needed post-editing. Focusing on the

[5] `https://pypi.org/project/google-trans-new/`

train set, we observe how the percentage for CNs is a bit higher than the HS one, which means that CNs needed more post-edition. The reason behind this could be the fact that CNs are naturally more complex than HS, as their objective is to counter an argument that someone else has stated, which results in poorer MT performance.

Regarding the test set, we observe that the percentage is also above 70%, but there is a slight difference from the train set: a total of 73 sentences were left without post-edition, as they were deemed to be too complex and needed to be rewritten. These sentences made up 5% of the sentences, so we could say that Train-CN and Test-CN had very similar percentages of post-edited/rewritten CNs.

4.2 Grammatical errors

Moving on to a more linguistic aspect of post-edition, grammatical errors should be mentioned, as they were rather abundant. Generally, most of these errors had something to do with verbs, either in terms of tenses, or in terms of conjugation. This might come as no surprise, given the morphosyntactic complexity of Basque verbs.

In the following examples we can see how sometimes, tenses in the original English sentence and the automatic translation did not match. Although sometimes these types of errors do not impact the overall meaning of the sentence (Example 1), some other times it is more apparent, as for instance in Example 2. Here the whole point of this CN is that they are using an event that happened in the past (the UK government asking Muslims to come help the country) as a counter argument for the HS. However, the MT in Basque is in the present tense, which implies that this action has not happened yet, which defeats the whole point of the CN.

The examples include the original CN (OG), as well as the MT and post-edited versions.

Example 1

OG *I am curious where you* ***get*** *those thoughts from.*

MT *Jakin-mina daukat nondik atera* ***zenituen*** *pentsamendu horiek.*

Example 2

OG *Then why* ***did*** *we ask them to come in the first place*

MT *Orduan, zergatik eskatu* ***diegu*** *lehenbailehen etortzeko?*

Post *Orduan, zergatik eskatu* ***genien*** *etortzeko, lehenik eta behin?*

In essence, we have observed that, as a general trend, the conjugation of verbs in Basque has been the cause of many errors, probably one of the most common of errors in post-edition. Some examples this phenomenon could be when the automatic translation used the verb 'dirudi' instead of 'iruditzen zaizu', or 'erakutsi' instead of 'erakuts dezakezu'.

Moving in to pronouns, English pronouns seem to have been problematic when translating them to Basque, as several erroneous translations were observed. It has to be noted

that English pronouns were translated to Basque in two ways: by using a pronoun in Basque (Examples 3 and 4), or through the Basque verb (Examples 5 and 6). In fact, Basque verbs are known to contain a lot of grammatical information within them, which is why it is possible to have elliptical pronouns. Whatever information the pronoun is giving, it is portrayed in the verb.

Many errors were found in both of these ways of translating pronouns: to begin with the pronouns, we see how pronouns like 'your' are translated to the Basque 'haien', while the English 'their' is often translated to 'zure'. In other words, the MT often mixes these two pronouns, using the third person pronoun 'haien' to translate 'your' (Example 3), while the second person pronoun 'zure' is used to translate the English pronoun 'their' (Example 4).

Example 3

OG *(...) by* ***your*** *logic the world would be a better place (...)*

MT *(...)* ***haien*** *logikaren arabera mundua leku hobea izango litzateke (...)*

Post *(...)* ***zure*** *logikaren arabera mundua leku hobea izango litzateke (...)*

Example 4

OG *(...) learning about a religion that many of* ***their*** *peers practice?*

MT *(...)* ***zure*** *ikaskide askok praktikatzen duten munduko erlijio bat ikastean*

Post *(...)* ***haien*** *ikaskide askok praktikatzen duten munduko erlijio bati buruz ikasteak?*

Moreover, there are some other examples where the pronoun difference in the original English and translated Basque is not as apparent: when the pronoun information is masked in the verb. In these examples, as previously mentioned, the two pronouns causing problems are 'you' and 'they'. In these cases, we can see how the English 'they' is translated to the Basque 'zu' (you) which is illustrated in verbs like 'baduzu' (Example 5). Not only that, but some instances were found where 'you' and 'their' where translated to the neutral singular pronoun 'hark' in verbs like 'duela' (Example 6).

Example 5

OG *If* ***they*** *love Sharia law so much, why don't they (...)*

MT *Sharia legea hainbeste maite* ***baduzu****, zergatik ez zara (...)*

Post *Sharia legea hainbeste maite* ***badute****, zergatik ez dira (...)*

Example 6

OG ***you*** *seem to be suggesting that (...)*

MT *(...) iradokitzen* ***duela*** *dirudi.*

Post *(...) iradokitzen* ***duzula*** *dirudi.*

4.3 Semantic related errors

Some semantic related incidents were also observed. For instance, there were some cases where the English word was translated to a seemingly correct word in Basque, but context showed a completely different sense of said word, resulting in a different Basque word. This is called polysemy, which is the "coexistence of many possible meanings for a word or phrase". This definition, together with the following definitions in this section have been extracted from Collins Dictionary [6].

This unsurprisingly created problems in translations, as a word in English with two different senses had two possible translations in Basque, one for each sense. The MT often did not identify these, mixing the Basque words and choosing a word with the incorrect meaning. For example, the word 'chill' in Example 7 is translated to 'hotza', as the sense of the word the MT has chosen is "to be cold but not frozen". However, in this context the word 'chill' has another meaning: "to relax, calm down", which in Basque is translated as 'lasai', as in the post-edited version.

Example 7

OG *This is manufactured outrage to create anti-Muslim sentiment.* ***Chill.***

MT *Hori musulmanen aurkako sentimendua sortzeko sortutako haserrea da.* ***Hotza.***

Post *Hori musulmanen aurkako sentimendua areagotzeko sortutako haserrea da.* ***Lasai.***

Another similar situation is Example 8: 'free from conflict' is translated to 'doan gataztatik'. The sense of the word 'free' that has been translated in the MT is "have it or use it without paying for it", which is in fact 'doan'. However, in this context the sense that we are looking for is more similar to "not restrained to", which would be translated as 'libre'.

Example 8

OG *(...) only [some countries] are* ***free from conflict.***

MT *(...) [herrialde batzuk] bakarrik daude* ***doan gatazkatik.***

Post *(...) [herrialde batzuk] bakarrik daude* ***gatazkatik libre.***

4.4 Hashtags

As previously mentioned, one of CONAN's main aim was to create a quality HS-CN dataset for CN generation, as the use of CNs has been deemed a better alternative to fight online hate. In other words, these HS-CN pairs were created in the context of the online world, and thus commonly include common characteristics of online texts, such as hashtags. When it comes to the MT data, hashtags created some problems, as the majority of them were

[6] `https://www.collinsdictionary.com/`

not translated into the target language. This created HS-CN pairs in Basque, but with the hashtags in English. Given the nature of the dataset, hashtags were deemed quite a significant part of the sentences, and it was decided to manually translate these hashtags in the post-edition.

Example 9

OG *That intolerance and hatred is one of the biggest problems our country faces* ***#spreadlovenothate.***

MT *Intolerantzia eta gorrotoa dela gure herrialdeak* ***#spreadlovenothate*** *duen arazo handienetako bat.*

Post *Intolerantzia eta gorrotoa hau da gure herrialdeak duen arazo handienetako bat* ***#maitasunazabalduezgorrotoa.***

4.5 Acronyms

As mentioned before, HS-CN pairs were collected with the goal of mimicking an online environment of these types of conversion. Therefore, the use of shortenings and acronyms comes as no surprise, such as expressions like 'tbh' or 'imho', or acronyms like 'fgm', or 'SME'. However, when it comes to translating them, many inconsistencies were found. For example, the expression 'imho' (in my humble opinion) was correctly translated to 'nire ustez', while in contrast, 'tbh' (to be honest) was not, as it was not even translated (Example 10). Both being shortenings, one was correctly translated, while the other needed to be corrected through post-editing.

Example 10

OG ***Tbh*** *I am struggling to find any Chocolate Eggs (...)*

MT *(...) txokolatezko arrautzak aurkitzeko borrokan nago.*

Post ***Egia esan,*** *(...) txokolatezko arrautzak aurkitzeko arazoak edukitzen ari naiz.*

When it comes to acronyms, some of them such as World War (WW1, WW2) and Small and Medium Enterprises (SME) were correctly translated into Basque. However, there was one rather problematic acronym, 'fgm', which stands for 'female genitalia mutilation'. This was wrongly translated to 'MGF' several times, which is in fact the abbreviation for the Spanish translation (Mutilación Genital Femenina) (Example 11). This is very interesting as it is the only instance in the 2000 post-edited pairs where part of the sentences were translated into another language that was not Basque or English. Finally, it is important to note that when FGM was in capital letters, some instances were found where it was correctly translated to the Basque 'emakumeen genitalen mutilazioa' (Example 12). Note how it was translated to the full sentence, not to the acronym.

Example 11

OG ***Fgm*** *is a serious human rights abuse.*

MT ***MGF*** *giza eskubideen gehiegikeria larria da.*

Post ***Emakumeen genitalen mutilazioa*** *giza eskubideen gehiegikeria larria da.*

Example 12

OG *(...) that this girl is at risk of* ***FGM****.*

MT *(...) haur honek* ***emakumezkoen genitalen mutilazioa*** *izateko arriskua duela.*

Post *(...) neska honek* ***emakumeen genitalen mutilazioa*** *izateko arriskua duela.*

This could imply that the automatic translation does a pretty good job with translating acronyms when they are common and have been traditionally used, but struggles when they are rather uncommon.

4.6 Sentence level errors

There were also some sentence-level errors found, especially in terms of duplicates and, in a fewer quantity, incomplete sentences. There were several cases, both in HS and CNs, where the sentence was translated to the target language, in this case Basque, but it was duplicated in a very specific format, which was consistent with all duplicates (see Example 13). The majority of the times they were perfect duplicates, although there were some cases where slight differences could be found, as for example some gender specifications, as in Example 14:

Example 13

OG *You sure?*

MT *['Ziur zaude', 'Ziur zaude']*

Post *Ziur zaude?*

Example 14

OG *Muslims have privileges that we do not have.*

MT *['Musulmanek guk ez ditugun pribilegioak dituzte', '****Emakume*** *musulmanek guk ez ditugun pribilegioak dituzte'.]*

Post *Musulmanek guk ez ditugun pribilegioak dituzte.*

Regarding incomplete sentences, it is rather common to find sentences where the main verb is missing, as in Example 15. Here, the MT version is missing the main verb 'do', 'egin', which is corrected in the post edition by adding the correct form of the verb, 'egiten dute'. It is a quite a peculiar error, as the English original is rather straightforward. One possible explanation is that the automatic translation mistook this instance as one where verb ellipsis is possible, when it is actually not.

Example 15

OG *It is a cultural thing, many Muslims* ***do not do it.***

MT *Gauza kulturala da, musulman askok* ***ez.***

Post *Gauza kulturala da, musulman askok* ***ez dute egiten.***

There is also an instance where half of the sentence was missing in the MT, as illustrated in Table 4: the totality of the second sentence is missing in the MT. The missing punctuation in the original CN (there is no full-stop or comma between 'people' and 'if'), which could have had an impact on the MT.

Original	**MT**	**Post-edited**
And yet most child rape, enslavement and murders are carried out by white **people if** you are serious about this stop scapegoating and let's deal with the real problem.	Eta, hala ere, haurren bortxaketa, esklabutza eta hilketa gehienak zuriek egiten dituzte.	Eta, hala ere, haurren bortxaketa, esklabutza eta hilketa gehienak zuriek egiten dituzte. **Hau serio hartzen baduzu, ez bota haiei errua eta konpondu dezagun arazo hau.**

Table 4: Example of an incomplete sentence

4.7 Typos in original English dataset

Several typos have been found in the original English version, which in a way make the dataset a little bit more realistic: if we are supposed to imitate online hate conversations, there will inevitably be some typos. Regarding the translation of those errors, it has been observed that sometimes the mistake in the original does not prevent the MT from being correct (Example 16).

Example 16

OG *(...)* ***Seond of all,*** *jihad, based on the concept of Muslim, (...)*

MT *(...)* ***Bigarrenik,*** *jihadak, musulmanen kontzeptuan oinarrituta, (...)*

However, it is rather unpredictable, as some other translations have been affected by typos. For instance, in Example 17, although it might not seem like it, 'whenever' is a typo, and if we look at the context, we will clearly see that the word that was supposed to be there was 'wherever'. This typo created an error in the MT version, which was corrected in the post-edition.

Example 17

OG *Muslim, as other human beings, have a right to live* ***whenever*** *they want.*

MT *Musulmanek, beste gizaki batzuek bezala,* ***nahi dutenean*** *bizitzeko eskubidea dute.*

Post *Musulmanek, beste gizakiek bezala,* ***nahi duten lekuan*** *bizitzeko eskubidea dute.*

4.8 Spanish post-edition

Apart from the Basque post-editing process, the Spanish data which was automatically translated by the Google API[7] was also post-edited by an intern student. In this case, the dataset in its totality underwent a post-editing process, as all HS-CN pairs in both train and test set were checked for correctness.

Regarding the percentage of post-edited HS-CN pairs, all results can be observed in Table 5. As we can see, the post-edited percentage in both train and test set is considerably lower than the ones previously mentioned for Basque (Table 3). In fact, out of all the CNs in the train set, only a 15% is post-edited, being this the least post-edited group in Spanish. Not only that, but the most post-edited group, the HSs in the train set, does not even reach the 24% of post-edited percentage. This is important to note, as these numbers are considerably lower than the post-editing percentages in Basque.

	Unique	**Post-edited**	**Percentage (%)**
Train-HS	383	90	23.50
Train-CN	4033	602	14.93
Test-HS	136	24	17.65
Test-CN	1268	256	20.17

Table 5: Spanish post-edition statistics

Qualitatively speaking, we observe that many of the errors present in the Basque post-edition were also found in the Spanish post-edition process. For instance, typos created problematic translations, as well as informal slang and acronyms or incomplete translations, which goes in line with Basque post-edition. Not only that, but semantically speaking, polysemy also created problems in Spanish translations, and some of the problematic words were the same for Spanish and Basque, as we can see in Example 18. This is the exact same sentence as in Example 7, but in Spanish, and we can see how the same error happened in the MT, regardless of the target language.

[7] `https://pypi.org/project/google-trans-new/`

Example 18

OG *This is manufactured outrage to create anti-Muslim sentiment.* ***Chill.***

MT *Esto es una indignación fabricada para crear un sentimiento anti-musulmán.* ***Frío.***

Post *Esto es una indignación fabricada para crear un sentimiento anti-musulmán.* ***Tranqui.***

Apart from these similarities, some post-editing issues were found in Spanish that were not found in Basque, as for instance gender concordance. As we can see in Table 6, the HS starts by talking about an issue surrounding women. Therefore, according to Spanish grammar, inflected words in the following sentences that make a reference to 'women' in the first sentence should agree in grammatical gender with it, i.e they should be in their feminine form. However, as we can see, the automatic translation fails to acknowledge this, and erroneously translates 'orgullosos' and 'verlos' to the masculine form, instead of 'orgullosas' and 'verlas'. This was not an issue in Basque, as it does not have grammatical gender.

Original HS	**MT HS**
Women of our culture that decide to become Islamic are so selfish. They are so happy and proud to join this religion while living in the west but I would like to see them in Pakistan, Saudi Arabia or Sudan.	Las mujeres de nuestra cultura que deciden volverse islámicas son muy egoístas. Están muy felices y **orgullosos** de unirse a esta religión mientras viven en el oeste, pero me gustaría **verlos** en Pakistán, Arabia Saudita o Sudán.

Table 6: Example of lack of gender concordance in Spanish

Summary To sum up, this section has informed about the Basque post-edition process, commenting on both quantitative statistics as well as the errors and tendencies found while editing. We have seen that, quantitatively speaking, the number of post-edited HS and CNs is considerably higher in Basque than in Spanish. However, when it comes to the actual errors, although some language specific errors were spotted, the majority of the post-editing instances were very similar in both languages.

5 Experimental Setting

This section will describe the experimental setting that was employed in this thesis, which includes the different versions of the datasets and its division, as well as the various experiments conducted with said datasets.

Datasets As mentioned in Section 3.1, the CONAN dataset (Chung et al., 2019) was used to fine tune several models in this thesis. We conducted the experiments by following the original division of the CONAN dataset into train and test sets, which can be seen in Table 7. As we can see, the whole corpus of HS-CN pairs consists of 6654 examples, from which 1288 were obtained in the original recollection, 2576 were obtained through a data augmentation process, and 2790 were obtained through the translation of French and Italian pairs. As we can see in Table 7, the augmented and translated pairs were used for training, while the original examples were used for testing.

	HS-CN count		
Original	1288	1288	Test set
Augmented	2576	5366	Train set
Translated	2790		

Table 7: Train-test split

However, there is one major difference between the original CONAN dataset and the ones used in this work: the language. While the original corpus contained information in English, Italian and French, our datasets contained information in Spanish and Basque. In order to obtain these novel version of the corpus, the whole original dataset was MT to both Spanish and Basque using the Google API [8]. These MT datasets were subsequently manually post-edited.

To sum up, all the experiments in this thesis were trained and tested using some form of the datasets above described:

1. **Two MT datasets**, in Spanish and in Basque.

2. **Two post-edited datasets**, in Spanish and in Basque.

3. **The original English dataset**, which was used to build baseline models

It is important to note that the total HS-CN pairs for training stated in Table 7 apply to all the datasets, with the exception of the post-edited Basque corpus: in this case, 2000 examples of the training set were post-edited, thus training the Basque post-edited experiments with fewer examples than the rest.

[8] `https://pypi.org/project/google-trans-new/`

Experiments Making use of these different versions of the CONAN dataset, two main types of experiments were carried out: (i) monolingual experiments in English, Spanish and Basque, which were then used as baselines for the crosslingual experiments (ii) crosslingual experiments, which included zero-shot and multilingual models.

Regarding monolingual models, 6 different models were trained, both with 5k and 2k training examples (thus, 12 in total): three Spanish models (es-mt, es-mt-post, es-post); and three Basque models (eu-mt, eu-mt-post, eu-post). Each of these three models in Spanish and Basque were trained with different combinations of the MT and post-edited datasets described above, as illustrated in Figure 4. An English model was additionally trained on the original language of the CONAN dataset, which was used as the Gold Standard (GS).

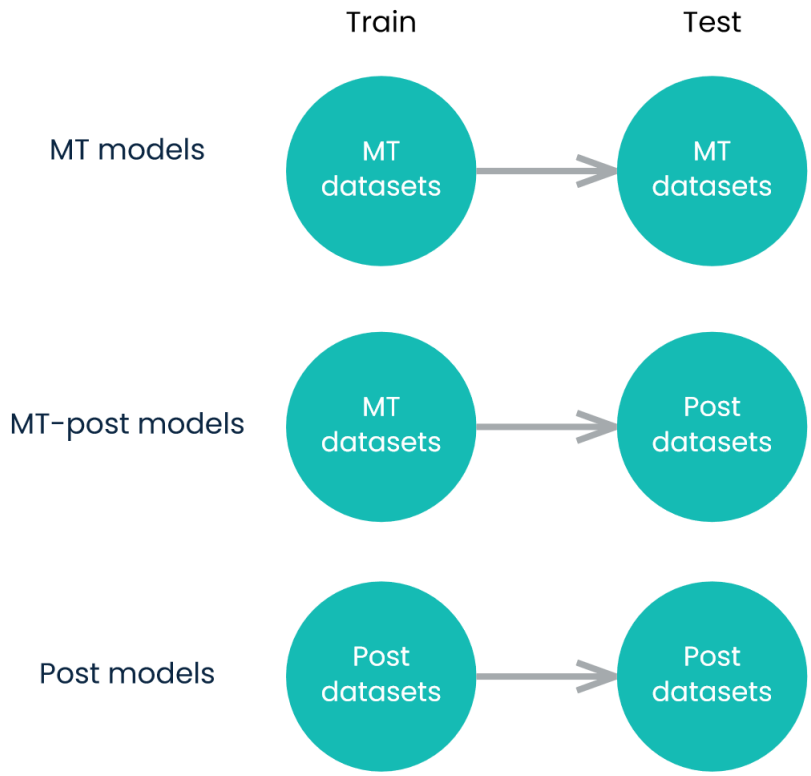

Figure 4: Dataset combinations for monolingual experiments (Spanish and Basque)

Regarding crosslingual experiments, both zero-shot and multilingual models were built, all using the post-edited datasets both for Spanish and Basque. To begin with, zero-shot models were trained with the aim of establishing the importance of having training data in the target language. In other words, we wanted to examine the feasibility of training a model in a language different from the generated output. In this regard, two models were trained (en2es and en2eu) where both the train and test sets included HS instances in English and CNs in the respective languages, either Spanish or Basque (Figure 5).

Regarding multilingual models, however, the aim was to check whether augmenting the data with other languages was beneficial or not. That is the reason why the following three experiments were carried out: all2en, all2es and all2eu, where the models were trained with the combined training data in English, Spanish and Basque, augmenting the training data to more than 12 thousand HS-CN examples. Both zero-shot and multilingual experiments are summarised in Figure 5.

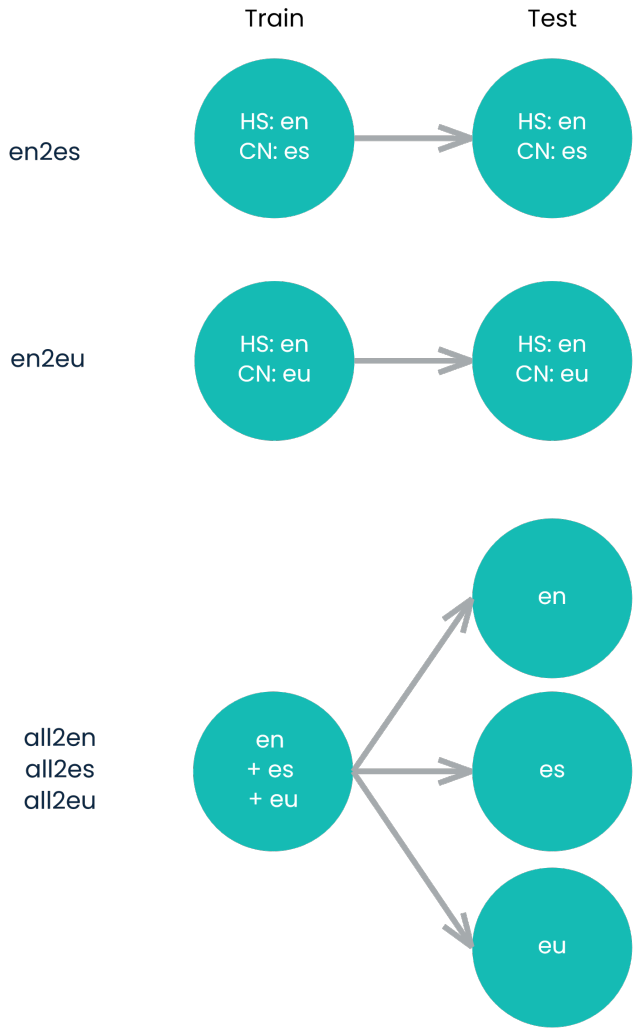

Figure 5: Dataset combinations for crosslingual experiments

The totality of these experiments above described were conducted using mT5, a multilingual language model described in Section 3.2. This model has been released in several different variants, such as mT5-small, mT5-base or mT5-large. In the initial stages of experiments, some models were trained with both mT5-small and mT5-base in order to see if the mT5 variant had an impact on the generated final text. After analysing these initial

results, mT5-base was chosen, as the output quality seemed slightly better and the training time was manageable. In terms of parameters, after trying various different configurations, it was decided to build these models with 50 epochs, a learning rate of 1e-3 and a batch size of 4, as this was the configuration that obtained the most satisfactory results.

Furthermore, shuffling HS-CN pairs both in the train and test set was considered, and several tests were carried out to check whether shuffling of the dataset had an effect on the output results. As the results did not have any noticeable differences when shuffled or not shuffled, it was decided that the experiments would be carried out with shuffled training sets and original-order test sets.

Finally, it needs to be highlighted that for the majority of the evaluation metrics, the generated output file was evaluated in its raw form. However, for some other evaluation metrics, namely RR and novelty, the generated output files were firstly preprocessed. More specifically, following Tekiroglu et al. (2020), all repeated instances were removed from the generated text files, obtaining the unique generated CNs. Not only that, but a shuffling procedure was also carried out for RR, in order to avoid similar CNs close together.

Summary To sum up, we section has the dataset used for the experiments, focusing on its division into train and tests sets, as well as its translation into Spanish and Basque and subsequent post-edition. Moreover, the configurations for the experiments have been described, both for monolingual and multilingual settings.

6 Empirical Results

The following section will present an in-depth analysis of both the monolingual and crosslingual experiments described in Section 5. It is divided in two sections: a quantitative evaluation, where the results of automatic metrics will be explored; and a qualitative evaluation, where a manual assessment is conducted and examined. As previously mentioned, the main objective of these experiments is to evaluate CN generation quality in Spanish and Basque, comparing the results of MT models with the post-edited models, as well as commenting on crosslingual results.

6.1 Quantitative Evaluation

This section informs about the quantitative evaluation that has been carried out, by commenting on several automatic metrics (BLEU, Rouge-L, RR and Novelty) that have been used to evaluate the performance of the trained models, both in terms of monolingual and crosslingual experiments.

6.1.1 Monolingual experiments

This section will focus on the comparison of monolingual models trained with MT and post-edited datasets, in Spanish and in Basque, thus focusing on answering RQ2: *In monolingual experiments, what are the differences in the generated output when models are trained with post-edited data, compared to Machine Translated (MT) data? Is it necessary to post-edit training data?*.

Not only that, but monolingual experiments were conducted twice, both with 5k and 2k training examples. Consequently, this section will also deal with RQ1: *In what ways does the training data size affect automatic metric results?*

The results for monolingual models can be observed in Table 8, both for 5k and 2k experiments. At first glance, we can observe how one specific model outperforms all the others: es-post, as its results, both for 5k and 2k models, are almost always the highest. One of the most interesting results was its considerably high BLEU score in 5k experiments (11.23), as it was almost 1.5 points higher than the next highest BLEU score, obtained by the English GS experiment (9.81).

However, it is important to note that there is one metric where es-post does not obtain the best results in, which is RR: it can be observed how both with 5k and 2k training examples, the English GS performs better than es-post in terms of repetition. Not only that, but when training the models with 5k examples, not only does the GS (5.73) obtain better RR results than es-post (6.32), but eu-mt also manages to get a slightly better result (6.29). Consequently, the high results in the n-gram overlapping metrics (BLEU and Rouge-L) might be influenced by its repetitiveness.

Moving on to its Basque counterpart, eu-post, we see a clear contrast to es-post: it obtains considerably lower results than the majority of the models. However, the fact the eu-post was only trained with 2k training examples needs to be considered, which is why its results will be further explored in the sections below.

	5k				2k			
Model	**BLEU**	**Rouge-L**	**RR**	**Novelty**	**BLEU**	**Rouge-L**	**RR**	**Novelty**
en (GS)	9.81	16.82	**5.73**	0.018	7.91	15.00	**7.76**	**0.048**
es-mt	8.08	16.13	8.99	0.003	7.05	15.23	13.83	0.002
es-mt-post	7.94	16.18	7.59	0.003	7.02	15.16	12.43	0.002
es-post	**11.23**	**18.99**	6.32	**0.042**	**8.29**	**16.52**	10.98	0.004
eu-mt	8.85	12.61	6.29	0.020	6.42	10.39	12.65	0.004
eu-mt-post	6.81	11.63	7.27	0.020	4.77	09.48	12.29	0.004
eu-post	-	-	-	-	5.12	10.25	14.88	0.006

Table 8: Monolingual results

Looking at the BLEU, Rouge-L and RR scores for the MT and MT-post models both in Basque and Spanish, we can see how MT models are generally higher than MT-post models: es-mt has generally higher results than es-mt-post, the same way eu-mt has higher scores than eu-mt-post. This is due to the overlapping nature of these metrics, thus the models trained and tested with the MT data (es-mt and eu-mt) are bound to have more overlapping to the training data than the models trained with the MT dataset, but tested with the post-edited test set (es-mt-post and eu-mt-post).

Finally, Table 8 also shows how crucial training size really is, as the results are clearly conditioned by the number of examples used in training: the results consistently deteriorate as the number of training examples decreases from 5k to 2k.

This phenomenon can be observed by comparing the results for eu-post (only trained with 2k examples) with the 5k Basque models, eu-mt and eu-mt-post. We can clearly observe how the 5k models are rather superior in all four of the metrics, specifically noting a very strong difference in RR scores: while eu-mt and eu-mt-post have rather low scores (6.29 and 7.27 respectively), eu-post obtains a considerably higher score of 14.88, which is roughly double what the 5k Basque models scored. These results imply that the effect of the training set size gets accentuated when it comes to RR, creating considerably more repetitive output as training size decreases.

Due to the importance of the training size pointed out in the paragraph above, the following analysis is divided in two parts: 5k models and 2k models. In each section, we will focus on the effect of the datasets used to train and test the models, i.e. whether they were trained with MT or post-edited data, or a mixture.

5K models The contrasts between model type results for 5k training examples can be observed in Figure 6 and 7. It is important to note that the Basque post-edited model

(eu-post), has only been trained with 2k examples, but has been included in the figures of both 5k and 2k models for comparison purposes.

By looking at the tendencies that Spanish and Basque models follow, Figure 6 shows how BLEU and Rouge-L follow a pattern: in the Spanish experiments, es-mt and es-mt-post have very similar scores, while the es-post scores are considerably higher. However, if we look at the Basque models, we can see how this tendency changes: we see a decreasing tendency, with eu-mt having the highest score, followed by eu-mt-post, and eu-post, which obtained the lowest score. This could imply that although post-edition is seemingly beneficial in terms of BLEU and Rouge-L in Spanish, eu-post (2k) fails surpass the performance of the MT 5k Basque models.

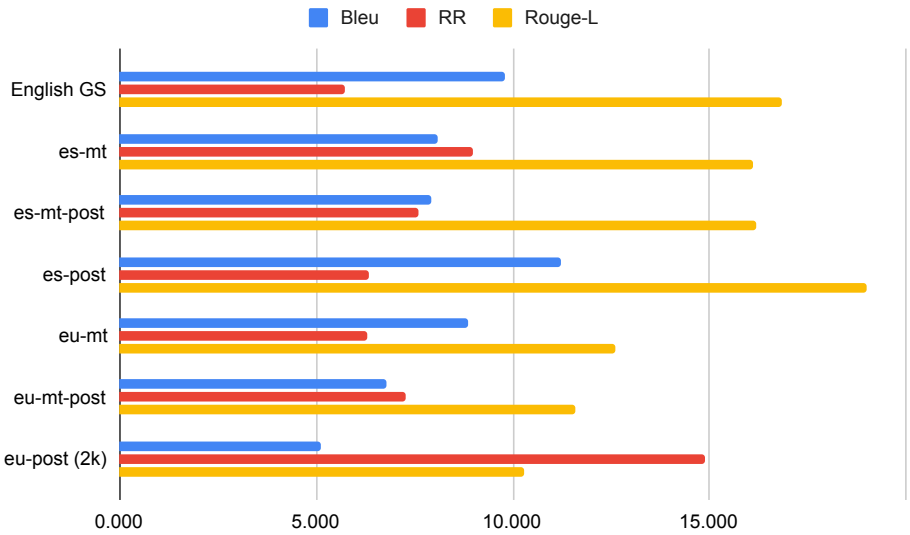

Figure 6: BLEU, Rouge-L and RR results with 5k training examples, except for eu-post

Observing the RR scores, we see a clear contrast between Basque and Spanish models: while the es-post RR score is the lowest followed by es-mt-post and es-mt, the Basque models show an opposite trend, eu-mt having the lowest score and eu-post the highest by far, as we can clearly observe an outlier in Figure 6. In other words, these results imply that in Spanish, training models with post-edited data creates seemingly less repetitive output, while with Basque the opposite happens: eu-post seems to create far more repetitive results than when training with MT data. As mentioned before, this could to be due to the effect of training size.

The last metric that needs to be mentioned regarding 5k experiments is Novelty (Figure 7). In general, although all results are very low, they seem to go in line with previous results: es-post obtains the highest score by far out of all the models, while eu-post performs the worst out of the Basque models.

Nevertheless, novelty scores for Spanish do deviate from previous BLEU, Rouge and RR results, as es-mt and es-mt-post novelty scores are lower than all Basque scores, even the

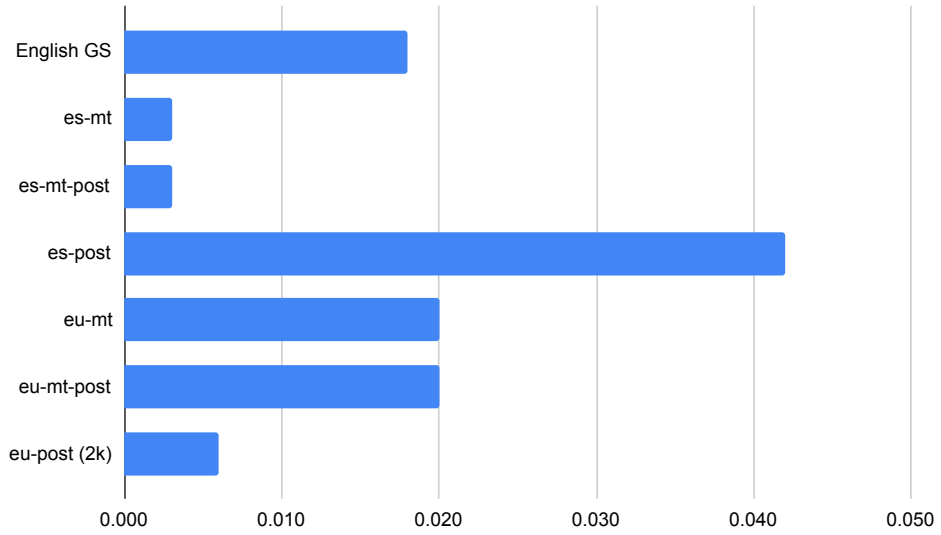

Figure 7: Novelty results with 5k training examples, except for eu-post

post-edited model, eu-post. This has never happened before, as eu-post has consistently obtained the lowest results.

These novelty results could imply that the output from the Spanish post-edited model is the most different one, the one which is creating the most different output when comparing it to the training data. This could either be a positive or negative phenomenon: it could mean that the output is in fact novel and of good quality, but it could also be possible that nonsense, low quality output has been created, obtaining a high novelty score. Therefore, correlation between novelty and manual qualitative evaluation is essential.

To sum up, 5k experiments have shown a clear tendency: while es-post consistently scores the highest results out of all the models, the opposite happens with eu-post, as it typically scores the lowest results. This implies that while post-editing has a clearly positive impact on Spanish CN generation results with 5k training examples, the same cannot be said regarding Basque experiments. The Basque post-edited model's poor results compared to the other Basque experiments could be because eu-post was trained with 2k examples; this will now be discussed by looking at the 2k metrics results.

2K models When it comes to BLEU, Rouge-L and RR, unlike 5k, there were no apparent outliers in terms of performance: as we can see in Figure 8 es-post does still have the highest BLEU and Rouge-L, and the second lowest RR score, but the difference with the other models has greatly decreased compared to 5k results (Figure 6). In other words, even though es-post still performs generally the best out of all models, the results across models have come closer, unlike with 5k experiments where es-post obtained considerably higher results.

In fact, this tendency also applies to Basque, as the results for eu-mt and eu-mt-post

are more similar with 2k trained models. This indicates the impact training size has on automatic metrics, as results start to deviate more from each other as the training size increases. This is especially notable in the RR scores in 5k and 2k experiments: as we can see in Table 8, the repetitiveness considerably decreases for all models from 2k to 5k experiments, as previously mentioned.

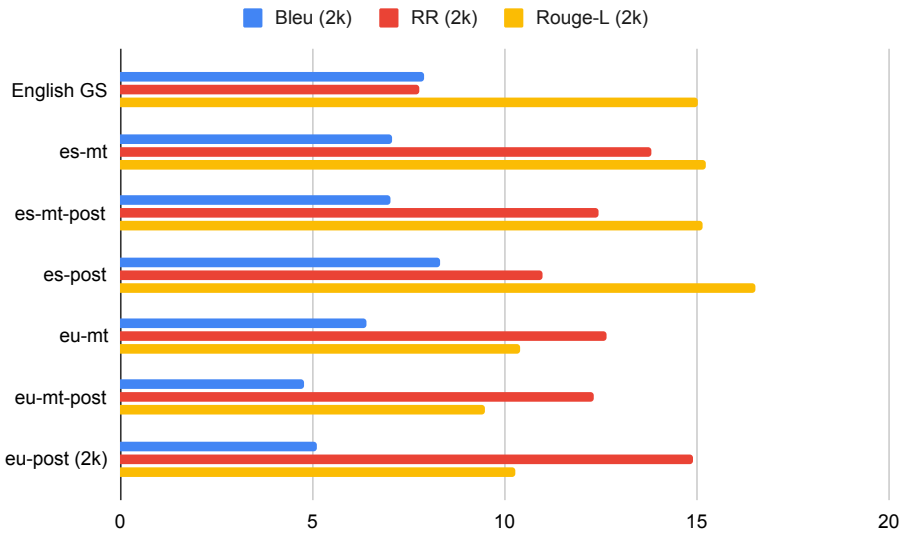

Figure 8: BLEU, Rouge-L and RR results with 2k training examples

Regarding BLEU and Rouge-L, we could say that the general trend seen in 5k experiments still stands here: es-post has the highest scores in Spanish, while eu-mt is the one with the highest performance in Basque. However, what is more interesting is the fact that eu-post does not have the lowest BLEU and Rouge-L scores anymore, as it performs slightly better than the eu-mt-post model, as well as being considerably closer to the eu-mt scores.

With regards to RR, we can see how the results show very similar results to those seen in 5k experiments for Spanish, as es-post still has the lowest score out the three Spanish models. Nevertheless, like BLEU and Rouge-L, Basque models do undergo a change in RR scores: eu-mt-post is the one scoring the lowest, closely followed by eu-mt and, in line with 5k experiments, eu-post keeps obtaining the highest score in RR.

However, it is important to note that eu-post's RR score is no longer an outlier: even though it is still the highest score, all other models seem to obtain very similar results, all above 10 points except for English GS (Table 8). Therefore, we can conclude that RR seems to be highly influenced by training examples, as scores in 5k experiments are considerably lower, leaving the only 2k model (eu-post) as an outlier (Figure 6), but when all models are trained with 2k examples, that contrast is no longer present (Figure 8). Not only that, but the English GS's score is also very interesting, as it consistently scores by far the lowest scores in RR, both in 5k and 2k.

When it comes to Novelty, there has been what seems a significant change in the 2k experiments: the Spanish post-edited score has dropped considerably, to the point where it is equal to eu-mt and eu-mt-post scores. Not only that, but in general we can observe how all Spanish models score equal or lower results than Basque models, which goes against general trend. It should also be highlighted that eu-post obtains the highest novelty score out of Spanish and Basque models, only being surpassed by the GS, which is something that has never happened before, as eu-post does not generally have the highest score in any metric.

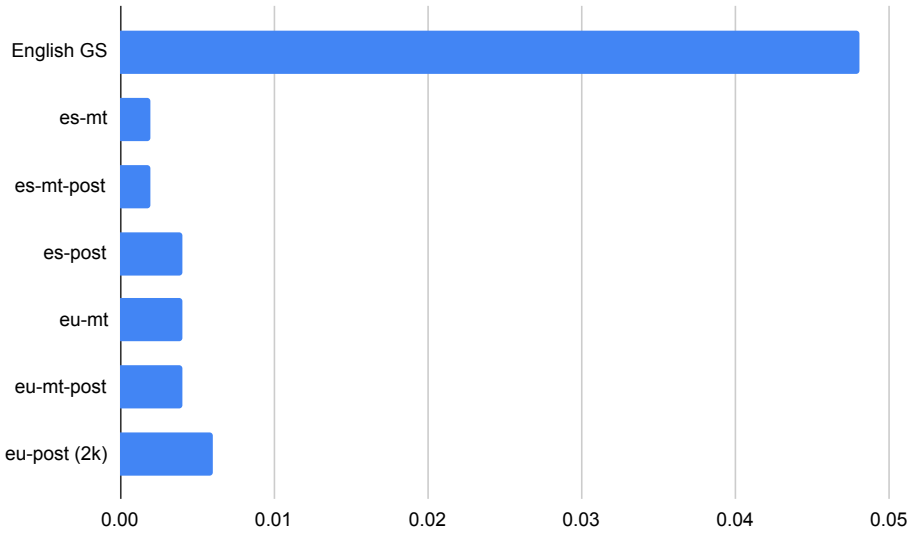

Figure 9: Novelty results with 2k training examples

After looking at novelty scores both in 5k and 2k experiments, we can say that a lot of inconsistencies were found in these scores, as there seems to be no specific trends that models follow (unlike the other metrics). This makes the novelty metric seem quite unpredictable, and poses the question of this metric being trustworthy enough. Due to cases like this one, a manual qualitative evaluation is essential in text generation tasks, in order to check for the correlation between automatic metrics and human measured quality, which will be discussed in 6.2.

To summarise, we have seen that the apparent dominance of es-post in 5k experiments is no longer present here: es-mt and es-mt-post get very close to es-post results, although not as much as to surpass it. In terms of Basque, we see promising results: eu-post does not obtain the worst results in all metrics anymore, as it has higher BLEU, Rouge-L than eu-mt-post, as well as higher novelty than both eu-mt and eu-mt-post. Not only that, but the eu-mt-post model obtains the lowest RR score. This could imply that with more post-edited training examples (i.e. 5k post-edited examples), Basque post-edited results could potentially improve and reach MT models' performance level, following the Spanish model's tendency.

Training size - Answer to RQ1 Results have indicated that the effect of training size is undeniable, as final results are often conditioned by number of training examples in many ways. In general, we have observed how results deteriorate as the training size decreases, which is especially notable in the RR scores. This comes as no surprise, as current text generation approaches such as language models (in our case, mT5) work best when more data is used to train them.

Not only that, but another trend was detected regarding training size: results seem to deviate considerably more from each other as training size increases. This again highlights the importance of having a considerable amount of training data, as it helps determine which configurations are more optimal than others.

MT vs. Post-edited - Answer to RQ2 Focusing on RQ2, specially the second part of the question (*Is it necessary to post-edit training data?*), it has been observed that Spanish and Basque seem to behave rather differently when it comes to models trained with MT or post-edited data. In Spanish, the experiments with the post-edited data, namely the es-post model, consistently obtained the most optimal results in the majority of the metrics, no matter the training size. This suggests the importance of post-editing data in Spanish, as MT data was unable to reach that level of performance.

With Basque, however, it was a rather complicated examination, as results presented greater variation. At first glance, results indicated a clear preference for the MT model (eu-mt), but 2k experiments created some ambiguity. While es-post obtained the best results both for 5k and 2k experiments, this was not replicated for Basque when we trained all models with 2k examples, as eu-mt obtained the best results, surpassing eu-post. However, eu-post obtained promising results, as it performed above the eu-mt-post model, and not extremely far away from eu-mt.

In other words, it seems like post-editing the data improves the output when compared with similarly sized training data, and thus training a Basque post-edited model with 5k examples could help considerably improve its performance. This should be further explored in the future.

All in all, it seems that, at least with the amount of data and type of configuration that we used, the Basque models trained with the MT dataset performed best at automatic metrics, while Spanish continually obtains the best results when the models are built with post-edited data. This will be further explored when automatic metrics are compared to the human evaluation in RQ5.

6.1.2 Crosslingual results

Moving on from the monolingual experiments, this section will explore the quantitative results of the crosslingual experiments. Two types of crosslingual experiments were conducted, the first being zero-shot models. By analysing these results, RQ3 will be studied: *In crosslingual environments, how do zero-shot models perform, compared to the monolingual baselines? Thus, how necessary is training data in the target language?.*

Not only that, but this crosslingual setting also involved several multilingual experiments, which will be analysed with the aim of answering RQ4: *How effective are data augmentation approaches such as monolingual MT datasets or crosslingual models, especially for low resource languages such as Basque?*

Due to the unpredictability of the Novelty scores in the monolingual results, the following analysis only focuses on the remaining three automatic metrics: BLEU, Rouge-L and RR.

The results for crosslingual experiments can be observed in Table 9, together with their corresponding monolingual baselines.

As we can see, there is a clear contrast between Spanish (en2es and all2es) and Basque (en2eu and all2eu) crosslingual models: results for BLEU and Rouge-L are considerably higher for Spanish models compared to Basque results: as we can see in Table 9, BLEU is above 10 for both en2es and all2es, while for Basque we obtained scores from as high as 5.12 to as low as 2.81. Regarding Rouge-L, Spanish models all score around 17-18, while Basque models show a lot more inconsistencies: although en2eu scores a very low Rouge-L (7.26), all2eu improves its score (11.11) but does not still reach the en2es or all2es scores.

	BLEU	**Rouge-L**	**RR**
en2es	10.03	17.78	**5.15**
en2eu	2.81	7.26	12.40
all2en	10.79	17.21	7.24
all2es	**11.36**	**18.83**	6.68
all2eu	6.46	11.11	10.92
Monolingual baselines			
en (GS)	9.81	16.82	5.73
es-post	11.23	18.99	6.32
eu-post	5.12	10.25	14.88

Table 9: Crosslingual results

Similarly, RR scores are consistently lower for Spanish models than for Basque, as en2eu and all2eu are the only ones with RR scores above 10, while en2es and all2es both score under 10, suggesting a rather big difference in terms of repetitiveness according to language: Spanish experiments produced less repetitive CNs compared to Basque models. These crosslingual results for BLEU, Rouge-L and RR go in line with monolingual trends, where Spanish models consistently obtained higher results than Basque models.

Focusing on a more specific analysis, some elements of crosslingual models and their baselines should be mentioned. By looking at the Spanish zero-shot results, we can conclude that the results for the zero-shot model (en2es) and the monolingual baseline (es-post) were very similar (Table 9). More specifically, the zero-shot results were slightly lower for BLEU (10.79) and Rouge-L (17.21) in the zero-shot experiment. Therefore, in general, we could

conclude that generating training data in the target language is a little more beneficial in Spanish than training it in English and generating the predictions in Spanish, although the difference is minimal.

Focusing on Basque zero-shot results, we can see a clear contrast from its Spanish counterpart: in this case, en2eu and its baseline have a rather big contrast when it comes to all three of the metrics. In fact, en2eu's BLEU and Rouge-L are incredibly low, not only compared to its baseline, but also compared to all other crosslingual results, as illustrated in Figure 10. Therefore, according to these results, having training data in Basque is essential, because training the models with English examples and generating the output in Basque produces rather poor results.

To finish with zero-shot models, RR results for both en2es and en2eu should be highlighted, as it does not seem to follow the same pattern that BLEU and Rouge-L follow: both in en2es and en2eu, RR scores seem to improve compared to their baselines, suggesting that the output created could be less repetitive. In other words, although metrics such as BLEU and Rouge-L do not seem to improve in zero-shot experiments, it seems like training the models in English creates less repetitiveness in the target language outputs.

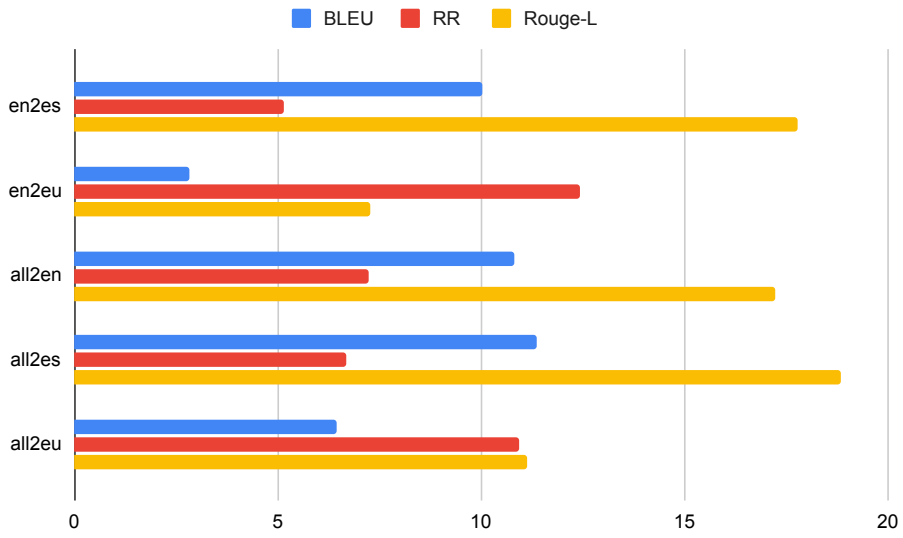

Figure 10: Crosslingual results for BLEU, Rouge-L and RR

With regards to the multilingual models, we can see that when we augmented the data by including the translations from the original English data as well as Basque data, there were virtually no differences in Spanish models: all2es and es-post have almost the same results in all three of the metrics. Regarding Basque, the results suggest that augmenting the data is actually beneficial: all2eu has a little higher scores than the baseline, which is interesting, as this means that the multilingual Basque model (all2eu) gets very similar results to eu-mt and eu-mt-post, specially in terms of Rouge-L. These are the monolingual models that generally obtained the highest results amongst Basque models.

Language of training data - Answer to RQ3 After this crosslingual analysis, the answer to RQ3 (*How necessary is training data in the target language?*) has been rather clearly stated: on the one hand, the difference between the zero-shot and baselines models in Spanish is minimal, with the monolingual baseline performing slightly better. Therefore, when it comes to Spanish, training models with the target language does not seem to be essential, as almost the same results can be obtained with the zero-shot model. On the other hand, Basque zero-shot models obtain incredibly low scores compared to the baseline, suggesting that training data in the target language is in fact essential for Basque.

Finally, the one metric that seems to behave similarly both in Spanish and Basque is RR, as both en2es and en2eu obtain lower scores than the monolingual baselines. This seems to indicate that zero-shot models are specially helpful when it comes to reducing the generated output repetitiveness.

Data augmentation - Answer to RQ4 In order to sum up the analysis of 6.1, RQ4 will be explored: *How effective are data augmentation approaches such as monolingual MT datasets or crosslingual models, especially for low resource languages such as Basque?* In short, we will examine how adequate these methods of data augmentation are in the specific task of CN generation.

In analysing this question, we have observed very different trends in terms of Spanish and Basque. To begin with, Spanish experiments indicated that MT data did not reach the same level of performance of the post-edited experiments, thus suggesting that augmenting the data through MT might not be a suitable approach for Spanish.

Not only that, but zero-shot experiments in Spanish show that the monolingual baseline performs slightly better than the zero-shot, concluding that data augmentation through a zero-shot approach is not essential. Moreover, multilingual models have shown virtually no differences in Spanish models.

Moving on to Basque results, automatic results indicate that, with the amount of data that we had available at the time of the experiments, models built with the post-edited datasets did not reach the performance of MT models. Consequently, according to these results using MT data as a data augmentation approach in monolingual environments could be beneficial for low resource languages like Basque.

Moving on to a crosslingual setting, as stated in RQ3, data augmentation through zero-shot approaches (i.e by using training data in other languages, in this case English) does not produce an acceptable output, as the obtained results are incredibly poor. Therefore, this method is not advised for Basque.

In terms of the multilingual setting, however, this changes: the generated output in Basque when trained with a multilingual dataset produces better scores than the monolingual baseline. This is probably due to the fact that the post-edited training data only contains 2k HS-CN pairs, so it greatly benefits from increased training examples. Based on this results, we could conclude that combining data in other languages with target language data could be a viable option when few data is available, as is the case of the majority of low resource languages, such as Basque.

6.2 Qualitative Evaluation

The following section will discuss the qualitative evaluation of the results, by commenting on the results of the manual evaluation. In doing so, the fifth and final research question will be explored: *Following previous studies which raised concerns regarding the lack of correlation between automatic metrics and human judgement in NLG, this thesis will assess the correlation between quantitative and qualitative evaluations.*

In preparation for the manual evaluation, six models were chosen: es-mt, es-post, en2es, eu-mt, eu-post and en2eu. After that, 20 HS-CN pairs were randomly chosen from the output of each model, so that they could be manually evaluated. Two annotators blindly evaluated all six of the models on five different criteria, on a scale from 1-5. Therefore, each annotator had to decide on a score between 1-5 for each of the 20 examples in all six models.

We based our qualitative evaluation on previous human evaluation criteria described in Chung et al. (2019) and Chung et al. (2020). These were the five criteria we chose:

1. Relatedness: it measures how related the CNs are with its corresponding HS, i.e. whether the CN is relevant given the HS that it is responding to.

2. Specificity: it states if the CN is rather generic or specific for the given HS it is responding to, thus replying to the question "can it be used for another completely different HS or not?"

3. Richness: in terms of language and vocabulary, it measures whether the CNs are simple or rather complex.

4. Coherence: it tells us whether the sentences make sense together, and if all ideas are clear and can be adequately understood.

5. Grammaticality: measures the grammatical correctness of the CNs.

Models	**Cohen's Kappa**
es-mt	0.7776
es-post	0.7742
en2es	0.8407
eu-mt	0.8751
eu-post	0.8502
en2eu	0.9054
Overall	0.8373

Table 10: Inter-annotator agreement (Cohen's kappa)

Once we had all the examples annotated, the average of each criteria was calculated, as well as the overall score for each model. This was done twice, as we had two annotators.

Once we had the results for both annotators, the inter-annotator agreement (IAA) was calculated, which can be observed in Table 10. It is worth noting that the agreement for Basque models was higher than for Spanish ones, in all of the models. Not only that, but if we look at the IAA of crosslingual models (en2es and en2eu) and compare it to their baselines, we see how crosslingual models consistently get higher results than their corresponding monolingual baselines (Table 10). Although some differences in IAA results were found, generally speaking we can see that the IAA was rather high in all of the models evaluated, as well as in the overall evaluation. This high IAA suggested that the annotation had been reliable, and thus the average between the two annotators' results was calculated, which can be observed in Table 11, together with the standard deviation for the overall score.

	Relatedness	**Specificity**	**Richness**	**Coherence**	**Grammar**	**Overall**
es-mt	2.85	2.65	3.33	3.65	3.63	3.22 ± 0.17
es-post	**3.61**	**3.31**	3.72	**4.25**	4.33	**3.84** ± 0.23
en2es	3.20	2.90	**3.83**	4.00	4.08	3.60 ± 0.04
eu-mt	2.88	2.65	2.85	3.13	3.05	2.91 ± 0.25
eu-post	2.45	2.05	3.40	3.95	**4.45**	3.26 ± 0.01
en2eu	1.85	1.53	3.70	3.98	4.43	3.10 ± 0.22

Table 11: Qualitative results (average from annotators)

6.2.1 Monolingual results

Focusing on the results for monolingual models (Figure 11), we can see how the five criteria used for evaluation can be divided in two groups: on the one hand, Relatedness and Specificity, which evaluate the quality of the generated CNs in relation to the HS they respond to; on the other hand, there is Richness, Coherence and Grammaticality, which solely take the generated CN into account for evaluation.

Focusing on the results for Relatedness and Specificity in Figure 11, we can see a clear pattern for Spanish and Basque models: in Spanish, es-post consistently gets higher results both in Relatedness and Specificity, with es-mt having poorer results. With Basque models, however, the opposite happens, as eu-mt performs higher than eu-post in both criteria. Consequently, these results suggest that, in Spanish, post-editing creates CNs that are more related to the HS they respond to as well as less generic responses; in Basque, however, post-editing does not seem to improve these two aspects, as the MT models consistently obtains higher scores, thus showing the lack of response quality the eu-post generated CNs have.

Moving on to the CN-specific criteria, Figure 11 shows how although some trends from the two previous criteria are maintained, some changes can also be observed. Regarding Richness, Coherence and Grammaticality, Spanish models follow the same trend, as es-post scores higher than es-mt in all three of them, clearly declaring itself to be superior to its automatic translation counterparts. Nevertheless, Basque models undergo a change from

the previous two criteria, as eu-post performs higher when it comes to these three CN-specific evaluations, specifically in Grammaticality: eu-post's score is considerably higher than that of eu-mt, even surpassing es-post.

Therefore, these qualitative evaluation of Basque experiments have shown that eu-mt performed the best at Relatedness and Specificity, while es-post excelled in Richness, Coherence or Grammaticality. This has resulted in eu-post scoring the overall highest qualitative score in Basque (Figure 11). This goes against quantitative results of Section 6.1, where the MT model consistently outperformed the post-edited experiments.

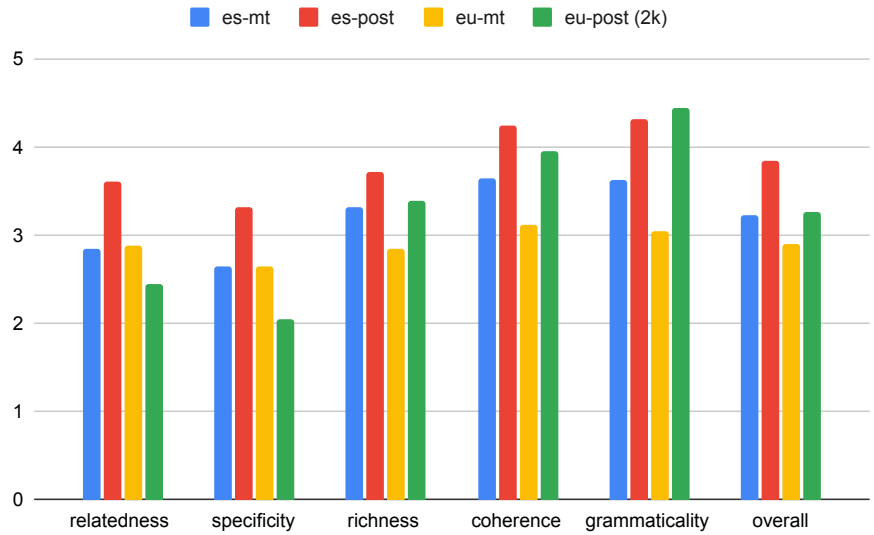

Figure 11: Monolingual qualitative results

In addition to this, assuming that the reason for high Specificity and Relatedness in eu-mt might be the increased training size, we could hypothesize that training eu-post with 5k examples could be a possible approach to improve eu-post's scores with regards to CN response quality.

To sum up the monolingual qualitative evaluation, we can see how, in line with previous results, es-post had the overall highest result, as it consistently obtained the highest scores. Not only that, but the two highest scores are from the two post-edited models (es-post and eu-post), suggesting that, qualitatively speaking, post-edition had a positive impact in both Spanish and Basque CN generation.

6.2.2 Crosslingual results

Regarding crosslingual Spanish models, the en2es zero-shot model was annotated in order for it to be compared to the Spanish monolingual baselines. As we can see in Figure 12, en2es is generally pretty consistent, in the sense that it usually performs lower than its

es-post baseline but higher than es-mt. There is only one exception, and that is Richness: it is the only criteria where en2es outperforms both es-post and es-mt.

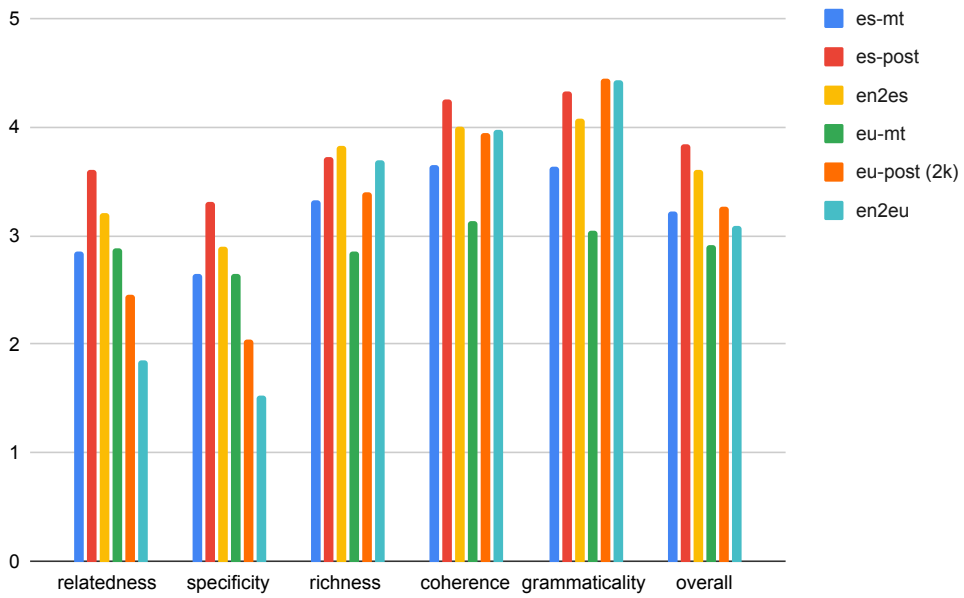

Figure 12: Crosslingual qualitative results and their monolingual baselines

Moving on to Basque, following the Spanish evaluation, the Basque zero-shot en2eu was annotated. In this case, we see a clear contrast between Specificity and Relatedness and CN-specific criteria, as we see that Relatedness and Specificity are considerably low compared to eu-mt and eu-post (Figure 12). In line with monolingual results, this changes when looking at the other three, as the results for en2eu considerably improves. In fact, en2eu's results for Coherence and Grammaticality are almost equal to eu-post, as well as surpassing es-post's score when it comes to Richness.

Therefore, the results for Richness seem to be quite significant when it comes to crosslingual models: both en2es and en2eu performed better than in other criteria, in fact, they scored the two highest scores, both surpassing their corresponding baselines. This suggests that zero-shot models improve the richness of the generated text in both Spanish and Basque.

Human correlation of automatic metrics - RQ5 This manual evaluation was carried out because several studies on CN generation have pointed out that automatic evaluation metrics often do not correlate well with human judgement (Pranesh et al., 2021; Qian et al., 2019). That is why RQ5 will be explored below: *Following previous studies which raised concerns regarding the lack of correlation between automatic metrics and human judgement in NLG, this thesis will assess the correlation between quantitative and qualitative evaluations*

After the analysis of the manual evaluation above, we have found that in the case of Spanish, quantitative and qualitative evaluations do in fact correlate: es-post consistently gets the highest results, suggesting that at least in Spanish, post-editing the data in fact improves the quality of the final output.

This correlation, however, is not present in Basque results. As previously mentioned, while eu-mt got the best results in the quantitative evaluation, the manual evaluation showed that the final output created by the eu-post model performed better in quality, specially in terms of Richness, Coherence and Grammaticality.

The contrast between Basque and Spanish results in terms of human correlation confirm the unpredictability of NLG automatic metrics, as quantitative results opted for the Spanish post-edited model but the Basque MT experiment (RQ1), while the qualitative evaluation opted for the post-edited models in both languages. This clearly emphasises the importance of a human evaluation step in these kinds of tasks.

7 Error Analysis

The following section will focus on particular errors found in the generated CNs, analysing them taking both the quantitative and qualitative evaluation into account, as well as the criteria used in the manual evaluation. In order to do so, Table 12 summarises the quantitative and qualitative results: both MT and post models of Spanish and Basque can be found here, as well as the zero-shot models for both languages.

Spanish To begin with, we can observe how quantitative and qualitative evaluations correlate in Spanish. Specifically looking at es-post, we see how it consistently gets the best results, both in the automatic evaluation, as well as the manual overall score. Not only that, but it is also the only model that scores above 3 for all five of the manual evaluation criteria, consequently suggesting a clear superiority with regards to the other models, both in terms of response quality and linguistic correctness.

	es-mt	**es-post**	**en2es**	**eu-mt**	**eu-post**	**en2eu**
BLEU	8.08	11.23	10.03	8.85	5.12	2.81
Rouge-L	16.13	18.99	17.78	12.61	10.25	7.26
RR	8.99	6.32	5.15	6.29	14.88	12.40
Manual overall	3.22	3.84	3.60	2.91	3.26	3.10
Detailed manual						
Relatedness	2.85	3.61	3.20	2.88	2.45	1.85
Specificity	2.65	3.31	2.90	2.65	2.05	1.53
Richness	3.33	3.72	3.83	2.85	3.40	3.70
Coherence	3.65	4.25	4.00	3.13	3.95	3.98
Grammatically	3.63	4.33	4.08	3.05	4.45	4.43

Table 12: Summary of qualitative and quantitative results

In fact, Example 19 shows an example of a CN generated by es-post, where the response has great quality in terms of all five of the criteria: the CN is related to the HS, as well as being specific, and it is linguistically rich, coherent and grammatical.

Example 19

HS *Si una niña tiene padres musulmanes, ¡corre el riesgo de sufrir mutilación genital femenina!*

CN *Muchos eruditos islámicos hacen campaña activamente contra la MGF, no se menciona en el Corán y está prohibido en varios países musulmanes. Este es un problema cultural, no religioso.*

The other two Spanish models, es-mt and en2es, are also very similar in the sense that both evaluations correlate, but none of them outperforms es-post: en2es closely follows its

results, while es-mt falls behind. This again implies that in Spanish, the post-edited model trained with 5k examples really improved the generated output in a rather consistent way, unlike Basque experiments, where some disparities were found.

Basque Moving on to the Basque experiments, we have previously stated that the correlation we observe in Spanish models, no longer can be found here: eu-post obtained rather low scores in the quantitative metrics, but it clearly obtained the highest overall result amongst the Basque models in the manual evaluation, especially regarding linguistic metric such as Richness, Coherence and Grammaticality.

In fact, one of the results that stands out in Table 12 is, as mentioned in the previous section, the high Grammaticality score of eu-post, especially compared to eu-mt. Eu-post, a model that has consistently obtained lower scores than eu-mt, suddenly improves its scores in the qualitative evaluation, obtaining an incredibly high Grammaticality score, by generating output CNs such as the one in Example 20:

Example 20 *[Koranaren aipua bakeari eta maitasunari buruz.] Koranak islamean sinesten ez duten pertsonei kalte egiteari buruz ezer esaten ez badu, nola da hori musulman guztiei buruzko hausnarketa?*

This suggests that post-edition really helps improve the linguistic correctness of the generated output, despite eu-post being trained with only 2k HS-CN examples. However, eu-post did have a rather apparent disadvantage: the two non-linguistic criteria in the qualitative evaluation (Relatedness and Specificity) obtained considerably lower scores than eu-mt (Table 12). As mentioned in Section 6.2, this might be due to the training size difference between eu-mt and eu-post, thus suggesting that the Relatedness and Specificity scores of eu-post could potentially improve if trained with more data.

An example of an unrelated and general CN generated by eu-post is illustrated in Example 21, as we can observe how the generated CN is not replying to the HS: while the HS is asking for the repatriation of Muslims, the CN is questioning why Muslims cannot build mosques in their country, which are two completely different topics. In short, although the actual CN is coherent and grammatical, it fails in being a good response to the HS.

Example 21

HS *Musulmanak gure herrialdetik atera beharko lirateke inork ez dituelako onartzen.*

CN *Zergatik ez zaie baimendu behar musulmanei meskita hemen eraikitzea? Herrialde honek kristautasunaz gain beste erlijio batzuk praktikatzeko askatasuna ematean sinesten zuela uste nuen.*

A very similar phenomenon happens if we look at en2eu's results, but rather more drastic. We can see in Table 12 that the quantitative results for en2eu were the worst scores out of all the models, even more so than eu-post, which did not perform good

itself. However, when we examine the manual evaluation overall score, we observe how, surprisingly, the en2eu score is rather good, being a little lower than eu-post but higher than eu-mt.

At first glance, it might seem like en2eu performs better than eu-mt, but a deeper and more specific analysis of the manual evaluation is essential: regarding CN-specific, language related criteria, en2eu did amazing, obtaining even better scores than eu-post in Richness and Coherence. However, it also performs extremely low when it comes to Specificity and Relatedness, creating a huge contrast between these two and the remaining three criteria. In short, it looks like the zero-shot model does help the output linguistically speaking, but fails in everything else, both in terms of automatic evaluation, as well as the response quality of the CNs, which is an important aspect of CNs. Example 22 shows an example of a low response quality CN generated by the en2eu model.

Example 22

HS *Since 2000 all suicide bombings were committed by a Muslim and no one handle this fact that about global terrorism.*

CN *Musulman britainiarrek beste edozein talde erlijiosok baino gehiago ematen diote karitateari per capita [esteka MEND-eko estatistiketara edo/eta meskita bateko ongintzako dohaintza handien adibidera], beraz, kosta egiten zait ikustea funts horiek biltzen diren meskitekin zer arazoa dagoen.*

In this example, we see how the generated response is neither related nor specific to the HS it is responding to: the hateful comment mentions terrorist attacks commited by Muslims, while the CN describes how Muslims give money to charity and build mosques, which does not respond to the HS.

Regarding Richness, crosslingual approaches and their performance in this criteria should be highlighted, as they obtained considerably high scores, not only in Basque but also in Spanish. In other words, augmenting the data with other languages improved the quality of the generated vocabulary. This is illustrated in the following example, where the generated Basque CN uses rather rich vocabulary:

Example 23

HS *Islam looks like a cult more than like a religion. It is beliefs make people rape children and commit violence. It should not be accepted in our country.*

CN *Islama ez da bateraezina zure bizimoduarekin, integraziorako oztopo handiena intolerantzia eta horrelako beldur faltsuak dira. Zenbat musulman ezagutu dituzu bortxatzaileak / hiltzaileak / esklaboak izan zirenak?*

Finally, the results for coherence also need to be mentioned, as it obtained really good results in virtually all the models, consequently becoming one of the most consistent criteria across all models, together with grammaticality (both these criteria were always above 3

points in all of the models). Therefore, all experiments seem to generate rather coherent CNs, including eu-post and en2eu, which failed in quantitative results, as well as Specificity and Relatedness. This is illustrated by the fact that very few incoherent sentences were found in the manual evaluation samples, one of which can be observed in Example 24.

Example 24

HS *Bortxatzaileen bandak musulmanak dira gehienak, Koranak agintzen du eta gure milaka nesken aurkako sexu indarkeria egiten dute. Nola jasan dezakegu hori?*

CN *Ez dut uste beste testu erlijioso batek ere horrelako zigorrik duenik. Zer esan nahi duzu?*

As we can observe, the generated CN is rather difficult to understand, both in terms of individually and, more importantly, in relation to the HS. In fact, this lack of coherence in the CN might be influenced by the fact that it is not neither related nor specific to the HS, so there is no clear context in which to interpret this CN. Instances where the generated CN had good Specificity and Relatedness and a lack of coherence were virtually non-existent.

Summary To sum up, this section has analysed the performance of both Spanish and Basque models, in terms of the specific generated CNs, and the manual criteria used to judge them. In order to do so, several examples have been provided to illustrate where the models excelled (Example 19, 20) as well as where there is still room for improvement (Example 21, 24).

All in all, these examples have shown that the majority of the times, the generated CNs in itself are acceptable, as they are generally coherent, grammatical and rather rich in terms of vocabulary. However, especially in Basque post-edited results, the generated CNs struggle to have good response quality, which could potentially be influenced by the low training size, and thus should be further explored.

8 Conclusions and Future Work

The aim of this thesis has been to analyse CN generation in Spanish and Basque. In order to do so, the CONAN dataset by Chung et al. (2019) has been used, which provided HS-CN pairs dealing with topics surrounding Islam and Muslims. For the purpose of this project, this dataset was MT into Spanish and Basque, as well as manually post-edited. With both the MT and post-edited datasets in Spanish and Basque, two types of experiment settings were explored: monolingual and crosslingual. These two settings were evaluated quantitatively as well as qualitatively.

To begin with, monolingual experiments were carried out both with 2k and 5k training examples. Answering RQ1, the findings of this analysis indicate the considerable impact of training size, especially in terms of RR. Additionally, the fact that results seem to deviate more from each other as training size increases was highlighted.

Furthermore, the quantitative evaluation of monolingual experiments especially focused on the difference between models trained with MT or post-edited data, in order to answer RQ2. Our findings suggested that both languages behaved rather differently: on the one hand, the Spanish post-edited dataset performed by far the best out of the other models, implying that post-edition is greatly beneficial in this case. For Basque, on the other hand, the model trained with the MT dataset performed the best. However, the effect of the training size is important to note: the result of the Basque post-edited model, which was only trained with 2k examples, seems to slightly improve when compared to similarly sized models. This leads us to believe that its results could potentially improve when trained with 5k HS-CN pairs.

Crosslingual experiments were also explored quantitatively. The results of the zero-shot experiments were described to answer RQ3, which showed contrast between Spanish and Basque results: while the Spanish results indicated that training data in the target language is not a necessity, Basque results suggested the opposite. Automatic metrics performed incredibly low on the Basque zero-shot experiments, suggesting that training data in Basque is essential.

Furthermore, the analysis of quantitative metrics was summarised in RQ4, by exploring the effectiveness of MT datasets or crosslingual settings as data augmentation approaches. The evidence from this examination suggests that, for Spanish, the model trained with the MT dataset did not reach post-edited performance. Moreover, both zero-shot and multilingual experiments showed virtually no differences to the monolingual baseline.

In terms of Basque, our findings have indicated that using MT datasets as a data augmentation approach in monolingual environments could be beneficial for low resource languages like Basque. With regards to crosslingual experiments, using data in a different language than the target language has been proved to decrease performance in Basque, but this changed when the Basque data was combined with data in other languages, i.e. in multilingual settings, as the model trained in this environment outperformed the Basque monolingual results. Therefore, we could conclude that augmenting the Basque data with data in other languages is beneficial when few data is available, as is the case of the majority

of low resource languages.

Moving on to the qualitative evaluation, a manual annotation process was conducted, where two annotators annotated selected models, evaluation five different criteria: Specificity, Relatedness, Coherence, Richness and Grammaticality. The findings of this evaluation emphasise the importance of a human evaluation in NLG tasks, as automatic metrics do not always correlate with human judgement (RQ5).

More specifically, our manual evaluation has shown that while the results for Spanish models correlate with automatic metrics, Basque experiments do not. Therefore, if we revisit RQ2, even though in Spanish we clearly see that post-edition helps improve the results, both in terms of automatic metrics and the manual evaluation, deciding on whether post-edition is worth it in Basque is rather complicated. This is because the Basque post-edited model performs rather low in the quantitative metrics, but the manual evaluation showed a lack of correlation with these results, as it obtained the highest manual overall out of all Basque models.

Finally, the effect of training size has also been examined in the manual evaluation, as it specially affected Specificity and Relatedness in the Basque post-edited model trained with 2k examples. It has been hypothesised that these two criteria are highly impacted by low training size, and thus could potentially improve with increased training examples.

This thesis concluded with an error analysis section, where the criteria used for manual evaluation was further studied and illustrated by examples of CNs generated in the previously examined experiments. This analysis highlighted that CNs on their own were generally of good quality, in terms of Coherence, Richness and Grammaticality. However, the majority of the generated CNs failed in the quality response to the HS they responded to, i.e. with regards to Specificity and Relatedness.

The work and analysis done on this thesis has provided this research field with several contributions. Firstly, the CONAN dataset by Chung et al. (2019) was MT and post-edited into Spanish and Basque, creating the first HS-CN pair dataset in these two languages. These resources were used to establish benchmark results for CN generation in Spanish and Basque, as well as introducing the first text generation work in Basque, to the best our knowledge. Moreover, the differences in using MT or post-edited in training were provided, as well as insights into crosslingual approaches in Spanish and Basque. Finally, appropriate data augmentation approaches were proposed depending on the target language.

After having worked on the different aspects of this thesis, some limitations have been found, which gives room for future possible work that would complement this one. For example, one of the main limitations has been the lack of language models in Basque, as there are rather few options. Consequently, the presence of better and improved language models for Basque generation would greatly benefit the task of CN generation.

Moreover, we have previously mentioned how the training size has greatly limited our analysis, especially in terms of the Basque post-edition. As a result, finishing the post-edition of all HS-CN pairs in the Basque training set would solve this problem in future analyses. In addition, this thesis has only worked with two different training sizes (2k and 5k), but it would be interesting to try different training sizes in future works.

References

Ahuir, V., Hurtado, L.-F., González, J. Á., and Segarra, E. (2021). Nasca and nases: Two monolingual pre-trained models for abstractive summarization in catalan and spanish. *Applied Sciences*, 11(21):9872.

Akhter, M. P., Jiangbin, Z., Naqvi, I. R., Abdelmajeed, M., and Sadiq, M. T. (2020). Automatic detection of offensive language for urdu and roman urdu. *IEEE Access*, 8:91213–91226.

Al-Hassan, A. and Al-Dossari, H. (2021). Detection of hate speech in arabic tweets using deep learning. *Multimedia Systems*, pages 1–12.

Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bai, X., Merenda, F., Zaghi, C., Caselli, T., and Nissim, M. (2018). Rug at germeval: Detecting offensive speech in german social media.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., Rosso, P., and Sanguinetti, M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63.

Benesch, S. (2014). Countering dangerous speech: New ideas for genocide prevention. *Available at SSRN 3686876*.

Benesch, S., Ruths, D., Dillon, K. P., Saleem, H. M., and Wright, L. (2016). Considerations for successful counterspeech. *Dangerous Speech Project*.

Bertoldi, N., Cettolo, M., and Federico, M. (2013). Cache-based online adaptation for machine translation enhanced computer assisted translation. In *MT-Summit*, pages 35–42.

Boer, V. d., Hildebrand, M., Aroyo, L., Leenheer, P. D., Dijkshoorn, C., Tesfa, B., and Schreiber, G. (2012). Nichesourcing: harnessing the power of crowds of experts. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 16–20.

Burnap, P. and Williams, M. L. (2016). Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data science*, 5:1–15.

Cao, Y., Wan, X., Yao, J., and Yu, D. (2020). Multisumm: Towards a unified model for multi-lingual abstractive summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):11–18.

Celikyilmaz, A., Clark, E., and Gao, J. (2021). Evaluation of text generation: A survey. *ArXiv*, abs/2006.14799.

Cettolo, M., Bertoldi, N., and Federico, M. (2014). The repetition rate of text as a predictor of the effectiveness of machine translation adaptation. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track*, pages 166–179.

Chung, Y., Kuzmenko, E., Tekiroglu, S. S., and Guerini, M. (2019). CONAN - counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. *CoRR*, abs/1910.03270.

Chung, Y.-L., Tekiroglu, S. S., and Guerini, M. (2020). Italian counter narrative generation to fight online hate speech. In *CLiC-it*.

Chung, Y.-L., Tekiroglu, S. S., and Guerini, M. (2021). Towards knowledge-grounded counter narrative generation for hate speech. *arXiv preprint arXiv:2106.11783*.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

De Gibert, O., Perez, N., García-Pablos, A., and Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

Dowlagar, S. and Mamidi, R. (2021). Hasocone@ fire-hasoc2020: Using bert and multilingual bert models for hate speech detection. *arXiv preprint arXiv:2101.09007*.

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., and Belding, E. (2018). Peer to peer hate: Hate speech instigators and their targets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Ernst, J., Schmitt, J. B., Rieger, D., Beier, A. K., Vorderer, P., Bente, G., and Roth, H.-J. (2017). Hate beneath the counter speech? a qualitative content analysis of user comments on youtube related to counter speech videos. *Journal for Deradicalization*, (10):1–49.

Esteban, A. and Lloret, E. (2017). Travelsum: A spanish summarization application focused on the tourism sector. *Procesamiento del Lenguaje Natural*, 59:159–162.

Fanton, M., Bonaldi, H., Tekiroglu, S. S., and Guerini, M. (2021). Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. *arXiv preprint arXiv:2107.08720*.

Faris, H., Aljarah, I., Habib, M., and Castillo, P. A. (2020). Hate speech detection using word embedding and deep learning in the arabic language context. In *ICPRAM*, pages 453–460.

Fersini, E., Nozza, D., and Rosso, P. (2018a). Overview of the evalita 2018 task on automatic misogyny identification (ami). *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:59.

Fersini, E., Rosso, P., and Anzovino, M. (2018b). Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

García-Díaz, J. A., Cánovas-García, M., Colomo-Palacios, R., and Valencia-García, R. (2021). Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings. *Future Generation Computer Systems*, 114:506–518.

García-Díaz, J. A., Jiménez-Zafra, S. M., García-Cumbreras, M. A., and Valencia-García, R. (2022). Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers. *Complex & Intelligent Systems*, pages 1–22.

Gitari, N. D., Zuping, Z., Damien, H., and Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Goenaga, I., Atutxa, A., Gojenola, K., Casillas, A., de Ilarraza, A. D., Ezeiza, N., Oronoz, M., Pérez, A., and Perez-de Viñaspre, O. (2018). Automatic misogyny identification using neural networks. In *IberEval@ SEPLN*, pages 249–254.

Goled, S. (2021). Why transformers are increasingly becoming as important as rnn and cnn? `https://analyticsindiamag.com/why-transformers-are-increasingly-becoming-as-important-as-rnn-and-cnn/`. Accessed: 31-05-2022.

Hammer, H. L. (2017). Automatic detection of hateful comments in online discussion. In *International Conference on Industrial Networks and Intelligent Systems*, pages 164–173. Springer.

Hasan, T., Bhattacharjee, A., Islam, M. S., Samin, K., Li, Y.-F., Kang, Y.-B., Rahman, M. S., and Shahriyar, R. (2021). Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *arXiv preprint arXiv:2106.13822*.

Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., and Taboada, M. (2020). The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, 4(2):155–190.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *ACL 2004*.

Markowitz, D. (2021). Transformers, explained: Understand the model behind gpt-3, bert, and t5. `https://daleonai.com/transformers-explained`. Accessed: 01-06-2022.

Mathew, B., Kumar, N., Goyal, P., Mukherjee, A., et al. (2018). Analyzing the hate and counter speech accounts on twitter. *arXiv preprint arXiv:1812.02712*.

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P., and Mukherjee, A. (2019). Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 369–380.

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., and Mukherjee, A. (2021). Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., and Yeung, D.-Y. (2019). Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049*.

Pamungkas, E. W., Cignarella, A. T., Basile, V., Patti, V., et al. (2018). 14-exlab@ unito for ami at ibereval2018: Exploiting lexical knowledge for detecting misogyny in english and spanish tweets. In *3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2018*, volume 2150, pages 234–241. CEUR-WS.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., and Camacho-Collados, M. (2019). Detecting and monitoring hate speech in twitter. *Sensors*, 19(21):4654.

Plaza-Del-Arco, F.-M., Molina-González, M. D., Ureña-López, L. A., and Martín-Valdivia, M. T. (2020). Detecting misogyny and xenophobia in spanish tweets using language technologies. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–19.

Plaza-del Arco, F. M., Molina-González, M. D., Urena-López, L. A., and Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, 166:114120.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Pranesh, R. R., Shekhar, A., and Kumar, A. (2021). Towards automatic online hate speech intervention generation using pretrained language model.

Qian, J., Bethke, A., Liu, Y., Belding, E., and Wang, W. Y. (2019). A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Reynolds, K., Kontostathis, A., and Edwards, L. (2011). Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops*, volume 2, pages 241–244. IEEE.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., and Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Sai, A. B., Mohankumar, A. K., and Khapra, M. M. (2020). A survey of evaluation metrics used for nlg systems. *ArXiv*, abs/2008.12009.

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., and Stranisci, M. (2018). An italian twitter corpus of hate speech against immigrants. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Schäfer, J. and Burtenshaw, B. (2019). Offence in dialogues: A corpus-based study. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1085–1093.

Schieb, C. and Preuss, M. (2016). Governing hate speech by means of counterspeech on facebook. In *66th ica annual conference, at fukuoka, japan*, pages 1–23.

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Silva, L., Mondal, M., Correa, D., Benevenuto, F., and Weber, I. (2016). Analyzing the targets of hate in online social media. In *Tenth international AAAI conference on web and social media*.

Sprugnoli, R., Menini, S., Tonelli, S., Filippo, O., and Piras, E. M. (2018). Creating a whatsapp dataset to study pre-teen cyberbullying. In *Second Workshop on Abusive Language Online (ALW2)*, pages 51–59. Association for Computational Linguistics.

Tekiroglu, S. S., Bonaldi, H., Fanton, M., and Guerini, M. (2022). Using pre-trained language models for producing counter narratives against hate speech: a comparative study. *arXiv preprint arXiv:2204.01440*.

Tekiroglu, S. S., Chung, Y.-L., and Guerini, M. (2020). Generating counter narratives against online hate speech: Data and strategies. In *ACL*.

Varab, D. and Schluter, N. (2021). Massivesumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vidgen, B. and Yasseri, T. (2020). Detecting weak and strong islamophobic hate speech on social media. *Journal of Information Technology & Politics*, 17(1):66–78.

Wang, K. and Wan, X. (2018). Sentigan: Generating sentimental texts via mixture adversarial networks. In *IJCAI*.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Wright, L., Ruths, D., Dillon, K. P., Saleem, H. M., and Benesch, S. (2017). Vectors for counterspeech on twitter. In *Proceedings of the first workshop on abusive language online*, pages 57–62.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2020). mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., and Çöltekin, Ç. (2020). Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.

Zhu, W. and Bhat, S. (2021). Generate, prune, select: A pipeline for counterspeech generation against online hate speech. *arXiv preprint arXiv:2106.01625*.