Master thesis in Intelligent Interactive Systems

Universitat Pompeu Fabra

Multilingual Lexical Simplification

Jorge S. Pimienta Castillo

Supervisor: Horacio Saggion

September 2021

Master thesis in Intelligent Interactive Systems

Universitat Pompeu Fabra

Multilingual Lexical Simplification

Jorge S. Pimienta Castillo

Supervisor: Horacio Saggion

September 2021

Contents

Dedication

I would like to dedicate this work to my friends and family, whose support was essential during the hardest times of this master's degree.

Acknowledgement

I would like to express my sincere gratitude to:

- My supervisor, **Horacio Saggion**, for his guidance, patience, dedication, and recognize the invaluable assistance that he provided during the development of this project.

- The team of NLP researchers: **Kim Cheng Sheang** for his advice in Machine Learning algorithms, and **Daniel Ferrés**, for his help and advise with BERT models. **Sanja Štajner**, for her recommendations to the evaluation of the model, and for offering a baseline and datasets to start with this research.

- All the annotators for their collaboration and participation in the review of the model performance.

Abstract

This report describes, implement, and evaluate one strategy for text simplification, namely, **Lexical Simplification**, that aims to reduce the complexity of some words in a sentence. This process is done in two main steps, the first, is a module that identifies the complex elements, and the second, is a module that replaces those elements for simpler variants.

For the first module, the system will use three different datasets that include human annotations in different languages: *English*, *Spanish*, and *German*, this will allow us to train a classifier that detects complex words.

For the second module, a pre-trained model for word prediction (BERT) will be used to generate the candidates, the candidates will be sorted based on Zipf's frequency, to later select the one with the highest value.

Finally, the complete system is evaluated using a test dataset, and a survey designed to collect human annotations and perception of *Fluency*, *Meaning* and *Simplicity*.

Keywords: *Complex Word Identification; Masked Language Model; Lexical Simplification, Word Frequency, Model Evaluation.*

Chapter 1

Introduction

Nowadays, information and communication technologies are very much present in teaching and learning models, as well as other areas, granting people access to a large amount of contents.

The field of Natural Language Processing (NLP), is a branch of artificial intelligence that studies how computers interpret human language, and it takes advantage of the fact those contents are available in digital format, this offers an opportunity to develop technological solutions that improve the way that information is delivered, and to enhance the readers and learners experience [1].

One of the major topics that are investigated in the NLP field is *Text Simplification*; this task aims to process, and interact with complex elements of a text, and transform them into an alternative version that contains similar information in a format that is easier to understand for the majority of the readers. There are two significant frameworks to approach the Text Simplification task. **Lexical Simplification** and **Syntactic Simplification** [2].

In the first one, Lexical Simplification, some words are detected, and then replaced for simpler variants, respecting their corresponding context, some of the main components of Lexical Simplification are the followings:

- Complex Word Identification (CWI): identifies which words are complex

- Word Generation: generate possible words for the identified complex words of the text.
- Evaluation of the candidate words

In the second approach, the Syntactic Simplification, the system will aim to replace the grammatical structure of the sentence to an alternative simpler version, some of the transformations may include: changing among different grammar tenses, changing between active and passive voices, and others [3]. Some of their main steps can be summarized as follows:

- Detection: Identify which portions of the text are complex
- Paraphrasing: Generate possible paraphrases to replace the identified complex portions of the text.
- Evaluation of the candidate paraphrases

This report will focus on the Lexical Simplification approach, and will study the main components that are involved in this task. At the end the results of the system performance will be presented, and also the evaluation of the system with human participants for different languages: *English*, *Spanish*, and *German*.

1.1 Motivation

Information can be presented with different levels of detail, abstraction, or emotions. This adds some uncertainty when someone needs to choose which word better expresses an idea, and forces the communication parties to trade-off between a simple or more elaborated (complex) communication.

Words and sentences can be complex for a number of reasons, such as: the grammar structure, number of words used, readability, and context, and culture, according to [4] and [5], because of this, approaches that focus in the replacement of the target word alone are blind to the context.

Let us consider the case of *dyslexia*, where an individual struggles to read and write certain configuration of words, exemplifies how learners can find an obstacle in their learning experience and knowledge acquisition, as shown in [6]. One way to overcome this problem is to make use of a *Lexical Simplification* system, as it can help by offering an alternative configuration of the problematic words.

Similarly, the case of Aphasia, a disorder that results from damage to some areas of the brain, that reduces the capacity to produce and understand language. In [7] a technique that involves *Syntactic Simplification* has been developed to modify the sentence configuration to make it easier for the reader to understand.

For those reasons, it is highlighted the importance of offering tools that solve these impediments, in order to offer an equitable and accessible learning framework [8].

1.2 Objectives

This project aims to complete the following objectives.

- Implement a trainable system that performs the task of **Complex Word Identification** (CWI), and report its performance for three different languages: English, Spanish and German.
- Implement a trainable system that performs the task of **Word Generation** for the words identified on the previous module, by offering context-aware candidates.
- Implement a module that replaces the identified complex word with another word from the candidate list, and evaluate its accuracy using human annotations for three different languages: English, Spanish and German.

Chapter 2

Methods

2.1 State-of-art

Research on the Lexical Simplification field has a dynamic trajectory, most of the theories are focused on uncovering the best ways to model language and manipulate their components. In this section, we will describe some of the works developed to address this challenge, by highlighting their most relevant characteristics.

A possible solution to the problem at hand is proposed in the **LEXenstein** framework [9], where it is described a Lexical Simplification pipeline. In their work the developers use candidates that are extracted from online dictionaries to replace those that were marked as complex. The structure followed by their system have been influential in the field because it describes the general framework taken to develop this type of systems, and is described in the following list:

- Complex Sentence
- Complex Word Identification
- Substitution Generation
- Simplified Sentence
- Substitution Ranking
- Substitution Selection

In other developments for Lexical Simplification like **LIGHT-LS** [10] we observe that is not necessary to always rely on simplified corpora like *Simple Wikipedia* and lexicons like WordNet [11] to train our models, for cases where those resources are not fully available, the solution proposed here encourages us to test with different data sources.

The **LexSiS Method** [12] presents a system that performs Lexical Simplification for the Spanish language, here the developers tackle this problem by proposing three main features, that are used to find the best candidate to replace a complex word, namely, a word vector model, word frequency, and word length.

In the **Complex Word Identification shared task of 2018** [13], a baseline to detect complex words is proposed using a logistic regression model as a classifier, this model uses two features described below:

- Word average length: that is the amount of characters of a word
- Average character length: that is the average amount of letters for a particular language, in the task: *English*, *Spanish*, and *German*

In this work, a dataset with human annotations has been created, this was obtained from native and non-native speakers for different languages. This dataset will later be used as the *golden labels* that feeds the logistic regression model during the training phase. This model later performs a binary classification, the output of this classifier can be interpreted as the probability of the text being "*complex*", or "*non-complex*". The results of this work are summarized in the following table:

Table 1: F1-Score for the baseline provided in the Complex Word Identification (CWI) Shared Task 2018 using different language model

Language	F1-Score
English	0.69
Spanish	0.72
German	0.79

Another approach is proposed in [14] for a multilingual complex word detection, this model uses a different type of features, for example, the *topic-relatedness*, and also a different type of dataset, here it uses Wikipedia articles for around 100 topics. This is done for three different languages; English, German, and Spanish [15]. In this work, the selection criteria of the candidates is done by computing the cosine similarity between the word-topic vector and the document. For the evaluation of the model it is proposed the F1-score as a precision measure because F1-score is a weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

In **EASSE: Easier Automatic Sentence Simplification Evaluation** [4], which is a python package that offers a standardized framework to facilitate Sentence Simplification systems, the F1-score is also used as a selection criteria.

Lexical Simplification also involves a step of transformation of the complex words detected in the previous step, this transformation may include some form of *Generation*, *Modification* or *Removal* of words. In 2018 Google developed a new language representation model called BERT [16]; which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layer, that can later be fine tuned to solve different tasks in the

Natural Language Processing field, such as text classification and text generation. For further details of Transformer Systems the reader is referred to [17] and [18]

BERT pre-trained models are available and distributed from different organizations [19], The pipeline tag **Fill-Mask** provides a framework that predicts the most probable word looking into to the preceding and subsequent context of a token in the sentence. It makes use of special tokens to indicate the beginning, end, and spaces to predict ("*fill*"). The prediction can generate one or multiple candidates that can be used to feed other modules to further tune the candidates selection.

In the image below, an example of **BERT** using the pipeline tag *Fill-Mask* is given, the input is a sentence with a **[MASK] token** and the model will predict the most probable word that fits in that context.

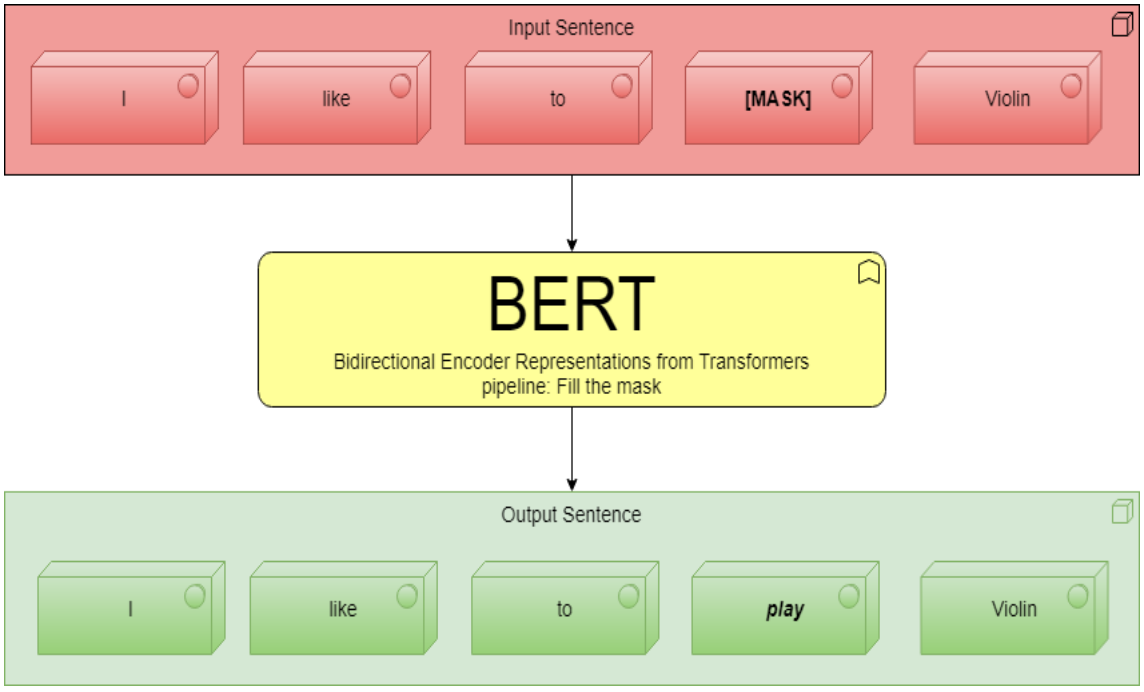

Figure 1: Example of BERT using pipeline: Fill-Mask on a sentence

Different pre-trained BERT models have been developed for multiple languages. Currently a multilingual model is available with more than 100 usable languages, and others trained for a specific language only. And for each specific language model there are variants trained with domain-specific corpus also available at [19], however, this is recommended mostly for high-resource languages, or languages where a large

corpora is available for training.

As shown in [20] the tasks where multilingual-BERT better performs are: **Name Entity Recognition** (NER), and **Part of Speech Tagging** (PoS), and this performance is even increased when the system is fine tunned for a specific context. In the following table, the results of NER and PoS for English, Spanish and German are reported:

Table 2: Multilingual BERT performance for the tasks of NER and PoS in English, Spanish and German [20]

Language	NER F1-Score	PoS Accuracy
English	90.70%	96.82%
Spanish	87.18%	96.71%
German	82%	93.99%

Language overlapping provides a bigger training input for the system, that allows it to go beyond memorizing a vocabulary for the task of **NER** and **PoS**, however, this acts differently for other tasks where grammatical ordering, and the linguistic structure differs [20]. Furthermore, in low-resource languages, which are those who lack large corpora, or manually crafted linguistic resources required for building statistical NLP application, multilingual BERT will not perform well [21].

Conversely, monolingual-BERT better performs for tasks related to text generation, namely, Question Answering, and Next Sentence/Word prediction, as it is analyzed in [22] using a *cloze test* where a percentage of the text are randomly masked and predicted back, this experiment used 15% of masked words, and their respective results for English and German are the following:

Table 3: *Cloze test* results for English and German int terms of how frequent the model provides the highest confidence score for the original subword, also known as, subword prediction accuracy. [22]

Language	Monolingual	Multilingual
English	45.92	33.94
German	43.93	28.10

This highlights BERT versatility to perform word's feature extraction, and word processing tasks. In the next sections, the implemented solutions will be further described.

2.2 Complex Word Identification (CWI)

2.2.1 Dataset

For this study, we used the data collected from the CWI Shared task 2018 [13], one dataset for each of the three target languages: *English*, *Spanish* and *German*.

The sentences in the datasets were extracted from Wikipedia articles and news, and the surrounding context was provided to the annotators for they to consider if that sentence could be considered complex for younger people, non-native speakers, or people with any cognitive impairments.

The procedures of handling the data followed the suggestions provided at the same repository [13], from that, the following components stand out: **the sentence** that contains a complex word, **the position** of the complex word within the sentence, and the **label** 0 or 1, to classify the word as not-complex or complex, respectively.

Take for example the sentence *"Both China and the Philippines flexed their muscles on Wednesday."*, The annotators provided the following annotations:

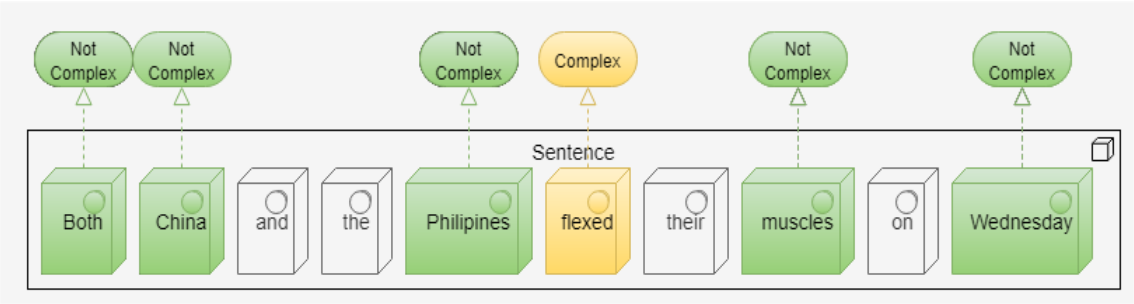

Figure 2: Sentence extracted from English Training dataset

In the example of the Figure, the target word **flexed** was marked as complex for at least 1 person. In some cases, n-grams can be also provided to the annotators. for some each language, the amount of participants is specified below:

- English: 20 annotators (10 native annotators, and 10 non-native annotators).
- Spanish: 10 annotators (a mix of native and non-native).

- German: 10 annotators (a mix of native and non-native).

Table 4: Binary and probabilistic golden labels from native and non-native annotators for English language

Target Word	**Annotators**	**Marked as Complex**	**Golden label**
flexed	20	8*	**1** or 0.4
China	20	0	**0** or 0.0
Philippines	20	0	**0** or 0.0
flexed their muscles	20	5**	**1** or 0.25
muscles	20	0	**0** or 0.0
Wednesday	20	0	**0** or 0.0

Note: *In the Table, the word **flexed** was marked as complex by 8 annotators (2 native and 6 non-native) the final binary label is 1, but there is also available as 0.4 for regression. **Similarly for the 3-gram *flexed their muscles* marked as complex by 5 annotators (3 native and 2 non-natives) or 0.25 for regression.

2.3 Classification task

2.3.1 Logistic Regression model

The *Logistic Regression model* is used most commonly in statistics to calculate the probability of occurrence of a binary dependent variable, in other words, models probability of output in terms of input [23]. In this project, this model is used to calculate if a word is complex or not given some **features** extracted from the target object, in this case: *the words* and the **golden labels** which come from the *human annotations* for each sentence. modeled as:

$$Logistic(\eta) = \frac{1}{1 + exp(-\eta)}$$

The implemented Logistic Regression model, uses the default regularization value of 1, a lbfgs (Limited-memory Broyden–Fletcher–Goldfarb–Shanno) solver which is robust to noise during the training phase [24]. The model supports L2 regularization, also known as Ridge Regression, and it will prevent overfitting due to the weights of the features in the model [25].

Finally, an adjusted value of max_iterations of 1000 was used, as the default value (100) was not enough for the model to converge in a solution.

2.3.2 Features selection

Originally the baseline included two features: *word length* (in characters) and *average word length per language*. The first one provides a simple approach to lexical complexity, as most of complex words are longer than their simplified versions, and the second, helps to mitigate cases where the language includes longer words on average when compared to others. For example, on average German words are longer (6.03 average word length) when compared to English (5.3 average word length). [14]

Later during the experiment, other two features were added: *frequency distribution*

of words given by a mathematical form known as **Zipf's law** and **Word Embedding** generated using BERT's neural network, that will be explained below:

The **Zipf's frequency** is a statistical model for human language that states that given some corpus of natural language, the frequency of any word is inversely proportional to its rank in the frequency table. Which is mathematically expressed as:

$$f(r) \propto \frac{1}{r^{\alpha}}$$

Where r is the rank that is assigned to every word in the text. For most texts, regardless of language, time of creation, genre of literature, its purpose, and others, and *f(r)* is its frequency in a natural corpus.

Since the actual observed frequency will depend on the size of the corpus examined, this law states frequencies proportionally: The foremost frequent word (r = 1) has a frequency proportional to 1, the second most frequent word (r = 2) has a frequency proportional to $1/2^{\alpha}$ (roughly twice as often), the third most frequent word has a frequency proportional to $1/3^{\alpha}$ (roughly three times more often), and so forth [26].

Word Embeddings are a learned representation, where words are mapped into a vector on a defined space [27], hence, words with similar meanings will be close in that embedding space. Some of the benefits of this representation are the following:

- They are numerical representations, therefore they can be used to train machine learning models.
- They rely on local statistics, so they can be used also in unsupervised learning models.
- They are context-aware.

BERT works by training a deep bidirectional representation from unlabeled text, it jointly conditions both: preceding and posterior content in all layers. Thanks to

this, BERT can be used as an input of a classifier to be fine-tuned in domain specific tasks, it allows to perform *knowledge transfer* to a different model, as shown in the following figure:

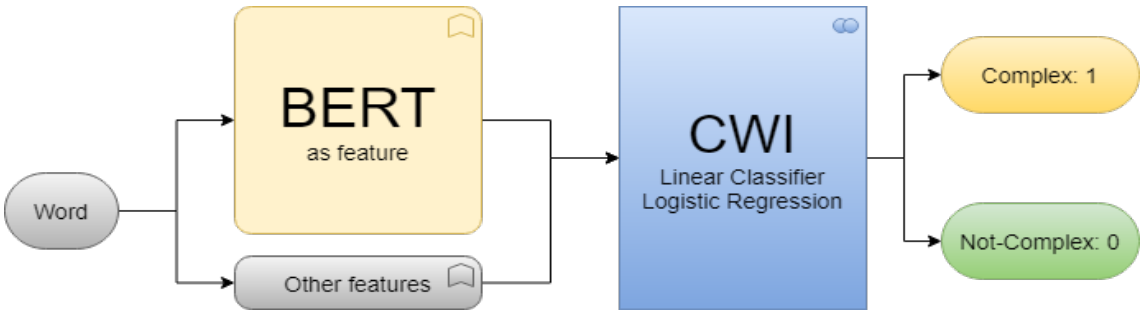

Figure 3: Complex Word Identification module

This approach provides a major computational benefits to pre-compute an expensive representation of the training data once and then run many experiments with less resource-consuming models on top of this representation [16].

2.4 Word generation

As it was introduced in previous sections, in addition to BERT's use as a feature extractor module, it can be used to predict a masked element on a string, by looking at the surrounding context. This deep bidirectional representation uses a *Masked Language Model* or **MLM**, and is trained by hiding some elements of the input words (usually 15% in the training corpus [16]).

The model requires the input to use a series of special tokens to indicate where the sentence begins, the separation between different sentences grouped in the same string, and the masked element that is required to be predicted with the model. The tokens may be placed in two ways, with the **tokenizer** method, and selecting the desired token, or by manually typing in string format **[CLS]** to indicate the beginning, **[SEP]** to indicate the sentence separation, and **MASK]** to indicate the token that we want to predict, as indicated below:

```
String "[CLS]" or {tokenizer.cls_token}
String "[MASK]" or {tokenizer.mask_token}
```

```
String "[SEP]" or {tokenizer.sep_token}
```

The context, is taken from the preceding and posteriors words to the target word in the sentence, this is illustrated in the next image:

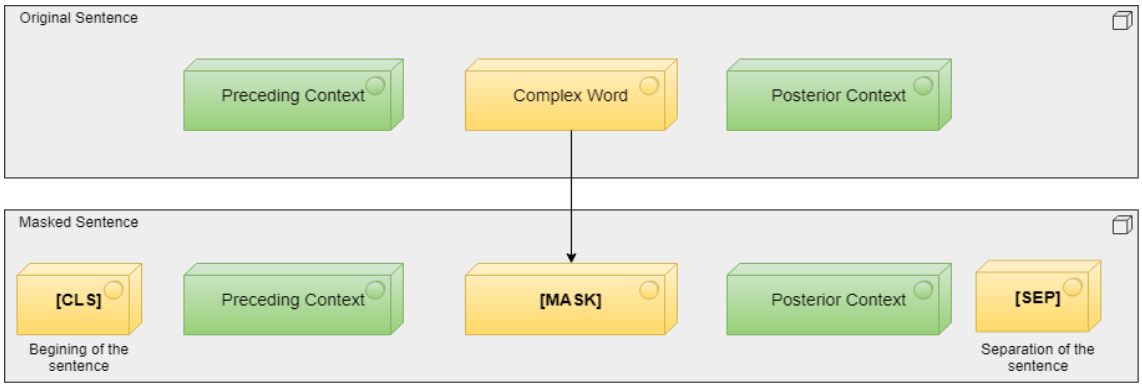

Figure 4: BERT Special tokens

Although this structure allows to generate a prediction based on context, the generated word will not necessarily be associated with the original word, because that same structure may be used in other more frequent scenarios.

To tell the model that the focus is on the word that has been hidden (masked token), it is necessary to follow the strategy defined in [28], in which the original phrase is duplicated and appended, the two phrases are separated using the special separation token (**["SEP]**") and the new set is used as input for the system. In the case of having multiple complex words, the procedure will be repeated recursively. The new structure will look as:

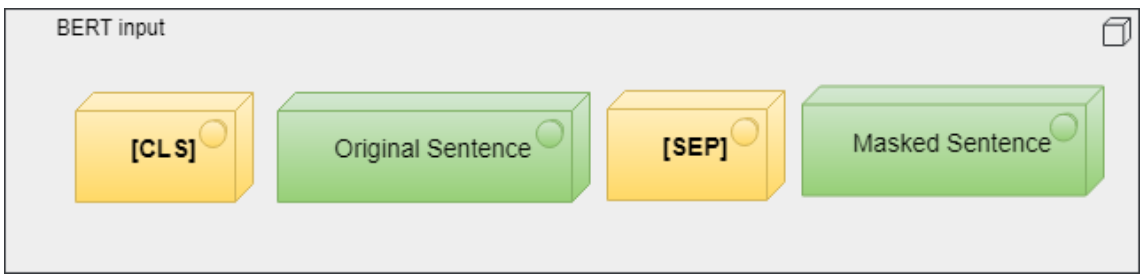

Figure 5: BERT target enforcement: Original sentence and Masked sentence

This approach allows the model to focus on words that will be closer to our target.

If it were not done this way, BERT would still be able to make predictions, but in cases where the context is poor it would generate unreliable results.

It is not in our interest to only generate predictions of a set, but to use BERT's language representations to find the best candidate to replace complex words given the surrounding context.

2.4.1 Language model selection

As explained in a previous section, monolingual BERT models better performs for text generation tasks [22], therefore, for this project different monolingual models was preferred over a multilingual model.

Monolingual models can be trained using a variety of text sources, furthermore, those sources can include one or multiple topics. For this project, it is preferred that the masked language models were trained with diverse topics, this is to cover different dimensions of human language. for this reason the following models were selected and imported into the system:

- **English**: bert-large-cased [16], this model was trained using *"BookCorpus"* [29] a dataset that contains 11,038 unpublished books and movie scripts, and also Wikipedia articles in English.
- **Spanish**: bert-base-spanish-wwm-cased [30], this model was trained using books, movie scripts, wikipedia articles in Spanish among other Spanish corpus. [31].
- **German**: bert-base-german-cased [32], this model was trained using Wikipedia articles in German, documents from *"OpenLegalData"* [33] and German news.

Each of the language models produce different candidates with a score that is sorted by the most probable word for each masked token. However, as we will explain in the next section, other features can be used to select the simplest among them.

2.4.2 Candidate generation

The amount of candidates that BERT can produce is variable, and can be adjusted each time BERT is executed, each candidate include the sentence with the decoded token, the score of that token, the token id, and the decoded token. The amount of shown candidates is given by the score value, where the highest comes first.

The structure of each candidate comes in a dictionary structure, and looks like the following example:

```
unmasker("This sentence contains a [MASK] token")
```

```
    {'sequence': 'This sentence contains a single token',
    'score': 0.05436881631612778,
    'token': 1423,
    'token_str': 'single'},
```

For this project five candidates will be generated per masked token, however a different selection criteria was additionally applied to the candidate list, for each item the Zipf's frequency [26] was computed, and sorted again to select the most frequent item [28]. This selection criteria offers a good approach to the objective of this study, because term familiarity is a valuable feature in simplifying text, as the most frequent words are most of the times the simplest among their alternative versions [34].

Chapter 3

Results

3.1 Complex Word Identification model evaluation

The metric used to evaluate the Complex Word Identification module was $F1-Score$ which is the harmonic mean of Precision and Recall, and gives an improved measure of the incorrectly classified cases than when using Accuracy alone.

Accuracy is often used when the **True Positives** and **True negatives** are more relevant, but the F1-score increases the effect of **False Negatives** and **False negatives**, this is important for this model because changing words that are not actually complex may reduce the simplicity instead.

The F1-Score is defined by the following formula:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

which is equivalent to:

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

Where **TP** is the total of True Positives, **FP** is the amount of *False Positives*, and **FN** or *False Negatives*

The F1-score is a relevant performance metric, specially for NLP applications be-

cause of the error correction of multi-step systems [35] like the one that is implemented on this project.

The Complex Word Identification module was tested with different feature combinations, and compared with the baseline value of the model; the results are presented in the next table:

Table 5: Model F1-score computed for English, Spanish and German

Language	**Baseline**	**Word frequency**	**BERT**	**Features combined**
English	0.69	0.71	0.81	0.81
Spanish	0.72	0.61	0.76	0.77
German	0.79	0.72	0.78	0.79

Overall, we observe that the model performance can be influenced by the combination of features, hence, choosing the right combination is important to ensure that the model will perform well, furthermore we found that each language benefits differently from each feature.

For example, when we focus in the **English** dataset, we see that adding the frequency of the word as a feature, an increase in system performance of 2 % is obtained, and when using the features extracted using BERT, an increase of 12 % in detection was obtained, combining the features derived in a similar performance.

Contrarily, in the **Spanish** dataset, the performance was reduced when the word frequency was included, and the best score was obtained from the BERT Word Embeddings.

Finally, for the **German** language, the new features did not provide a significant improvement to the performance of the Complex Word Detection for that language.

3.2 Word generation evaluation

As it was described in previous sections, the output of the Candidate Generation module is a list of words that is sorted based on their frequency value. To evaluate if this selection provides an appropriate approach to lexical simplification, it was required to evaluate the model with the help of human annotators.

The annotators were given a survey that contained 10 pairs of sentences, the original sentence extracted from the test dataset using a random selection, and a simplified version extracted from the output of the simplification model is described in this report.

They were asked to answer questions related to the generated sentence, and to compare its level of fluency or grammatical correctness,meaning or preservation of the original amount of information, and simplicity as these parameters provide relevant information about the performance of the system. Methods were based on previous experiments [36], this proceeded in the following questions:

1. **Fluency**: Is the second sentence grammatically correct?
2. **Meaning**: How much does the second sentence express the same content of the first sentence?
3. **Simplicity**: In which degree the second sentence is easier to read and understand?

To investigate this statistically, all questions were ranked using the **Likert Scale** [37]; The main advantage of the Likert Scale is that they it is a well-known method for survey collection, therefore it is easily understood from the participants, for this, each sentence will be scored from "1" which represents a low value, to "5" to represent a high value.

This survey was developed and applied in a similar fashion for our three target languages; English, Spanish and German. Their results are summarized in the next sections.

3.2.1 English language system evaluation

In the English survey a total of **30 participants** provided their feedback. The participants are a mix of native and non-native speakers, and the majority of the participants had a high level of language proficiency, i.e. Superior to C1, as shown below:

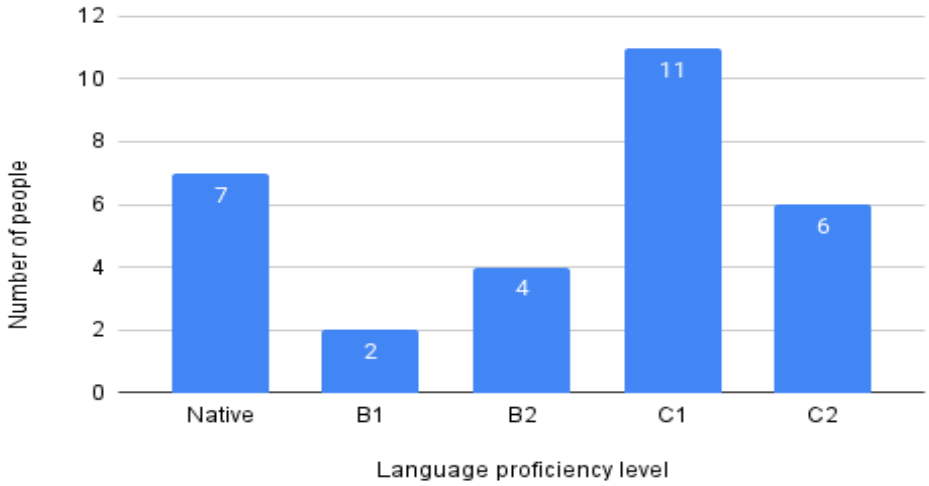

Figure 6: English proficiency level

The participants also belonged to a mix of gender groups, from which the majority of participants had completed Bachelor's or equivalent and superior educational study level, as summarized in the following graphs:

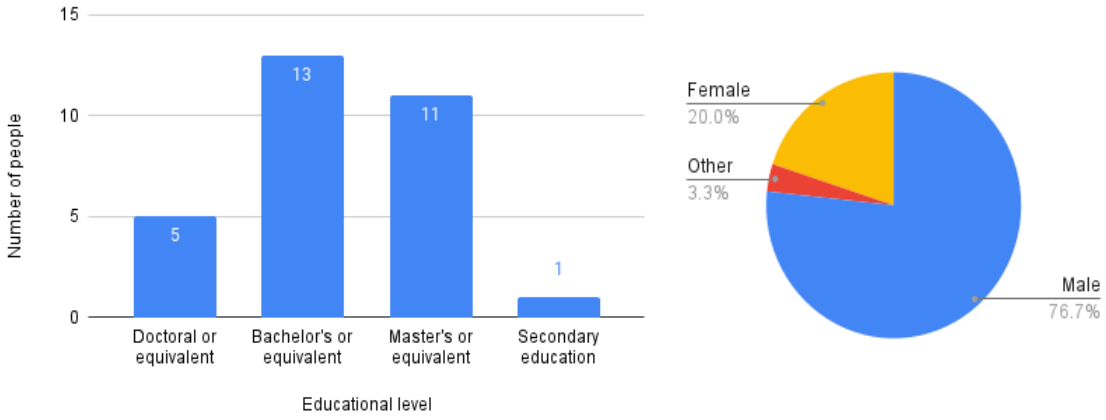

Figure 7: Educational level and gender distribution

Now we present some of the sentences that the participants were asked to rank, and their respective scores. The full list of sentences is attached in the Appendix in the English section.

Example 3.

Original Sentence: "#5-11 His government stands accused by Human Rights Watch of not taking **adequate** measures to protect the **nation's citizens**."

Simplified Sentence: "#5-11 His government stands accused by Human Rights Watch of not taking **enough** measures to protect the **American citizens**."

Example 7.

Original Sentence: "It's a complete **tragedy** for the town."

Simplified Sentence: "It's a complete **loss** for the town."

Table 6: Evaluation results for the English Examples: 3. and 7. using the Likert scale [37] from 1 to 5

Parameter	Example 3.	Example 7.
Fluency	4.52	4.66
Meaning	3.97	3.59
Simplicity	3.31	3.52

The score for Fluency, Meaning and simplicity of all the sentences is summarized in the next graph:

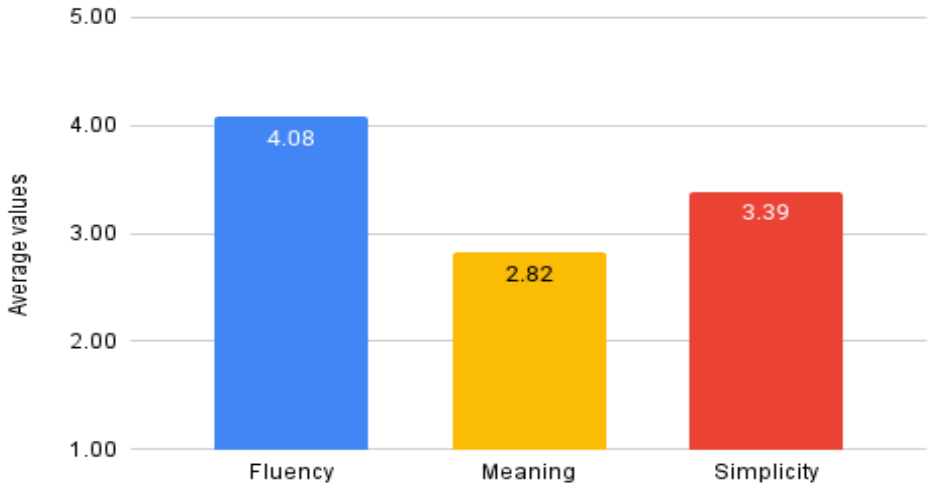

Figure 8: Average score of the 10 pairs of sentences for the values of Fluency, Meaning and Simplicity.

Overall we observe that for the English language the highest value corresponds to the Fluency parameter, suggesting that the model is capable of expressing sentences in *Natural* way with a score of **4.08**. The Simplicity score is **3.39** indicating how much the users perceived the sentence as more simple than the original. And finally the participants gave a score of **2.82** for how much meaning was preserved during the sentence transformation.

3.2.2 Spanish language system Evaluation

In the Spanish survey a total of **47 participants** provided their feedback. The participants are also a mix of native and non-native speakers, and similarly to the previous group, the majority of the participants had a high level of language proficiency, i.e. Superior to C1, as shown below:

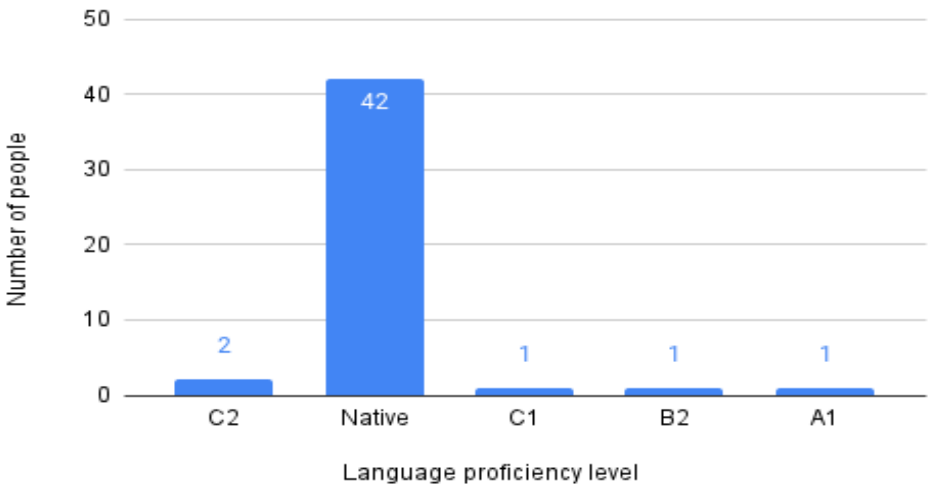

Figure 9: Spanish proficiency level

The participants also belonged to a mix of gender groups, from which the majority of participants had completed Bachelor's or equivalent and superior educational study level, as summarized in the following graphs:

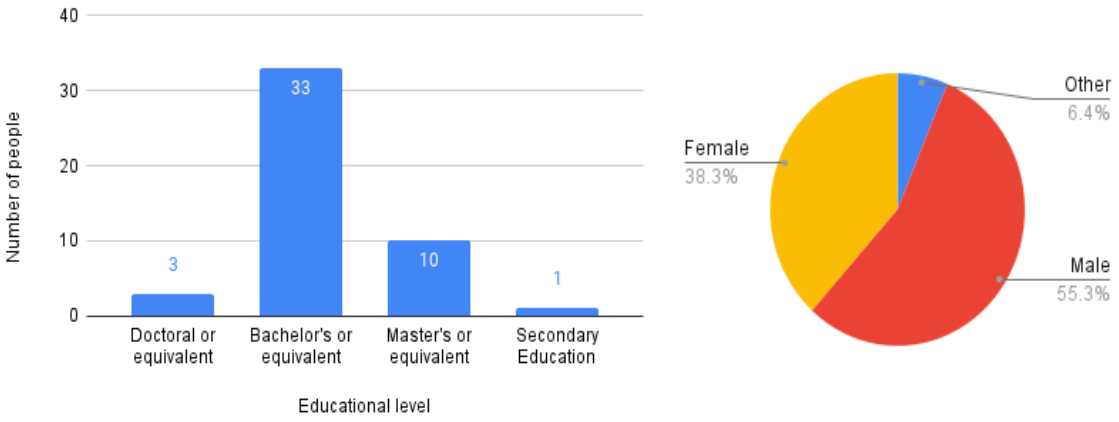

Figure 10: Educational level and gender distribution

Now We will also present some of the sentences that the participants were asked

to rank (here translated to English), and their respective scores. The full list of sentences is attached in the Appendix in the Spanish section.

Example 2.
Original Sentence: "Mokujin es un muñeco hecho de madera de roble de hace 2000 años, de los bosques antiguos de Japón, donde se **especulaba** la presencia de algo mágico."

Original Sentence (English): "Mokujin is a doll made of oak wood from 2000 years ago, from the ancient forests of Japan, where the presence of something magical was **speculated**."

Simplified Sentence: "Mokujin es un muñeco hecho de madera de roble de hace 2000 años, de los bosques antiguos de Japón, donde se **creía** la presencia de algo mágico."

Simplified Sentence (English): "Mokujin is a doll made of 2000-year-old oak wood, from the ancient forests of Japan, where the presence of something magical was **believed**."

Example 8.
Original Sentence: "Las dos fueron erigidas y **consagradas** al mismo tiempo: la de Santa María fue consagrada un día más tarde que la de San Clemente, en el año 1123."

Original Sentence (English): "Both were erected and **consecrated** at the same time: that of Santa María was consecrated a day later than that of San Clemente, in the year 1123."

Simplified Sentence: "Las dos fueron erigidas y **dedicadas** al mismo tiempo: la de Santa María fue consagrada un día más tarde que la de San Clemente, en el año 1123."

Simplified Sentence (English): "Both were erected and **dedicated** at the same time:

that of Santa María was consecrated a day later than that of San Clemente, in the year 1123. "

Table 7: Evaluation results for the Spanish Examples: 2. and 8. using the Likert scale [37] from 1 to 5

Parameter	Example 2.	Example 8.
Fluency	3.98	3.79
Meaning	3.90	3.55
Simplicity	3.75	3.64

The score for Fluency, Meaning and simplicity of all the sentences is summarized in the next graph:

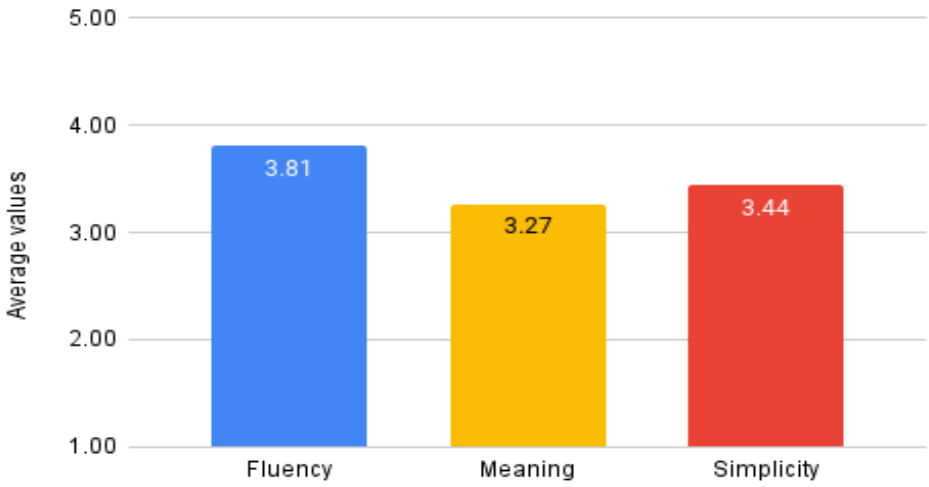

Figure 11: Average score of the 10 pairs of sentences for the values of Fluency, Meaning and Simplicity.

Overall we observe that for the Spanish language the highest value corresponds to the Fluency parameter with a score of **3.81**. The Simplicity score is **3.44** And finally a score of **3.26** for how much meaning was preserved during the sentence transformation.

3.2.3 German language system evaluation

In the German survey a total of **11 participants** provided their feedback. The participants are a mix of native and non-native speakers, with a more spread levels, namely from A2 to Native. as shown below:

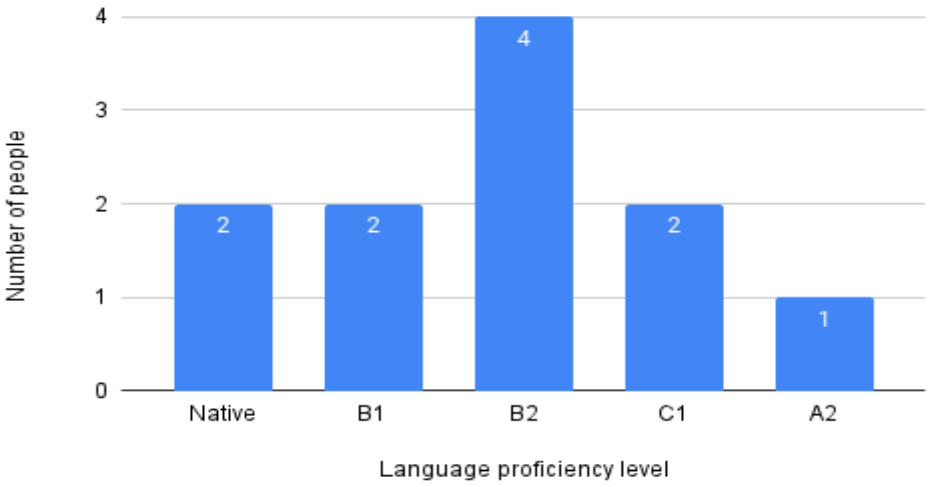

Figure 12: German proficiency level

The participants also belonged to a mix of gender groups, from which the majority of participants had completed Bachelor's or equivalent and superior educational study level, as summarized in the following graphs:

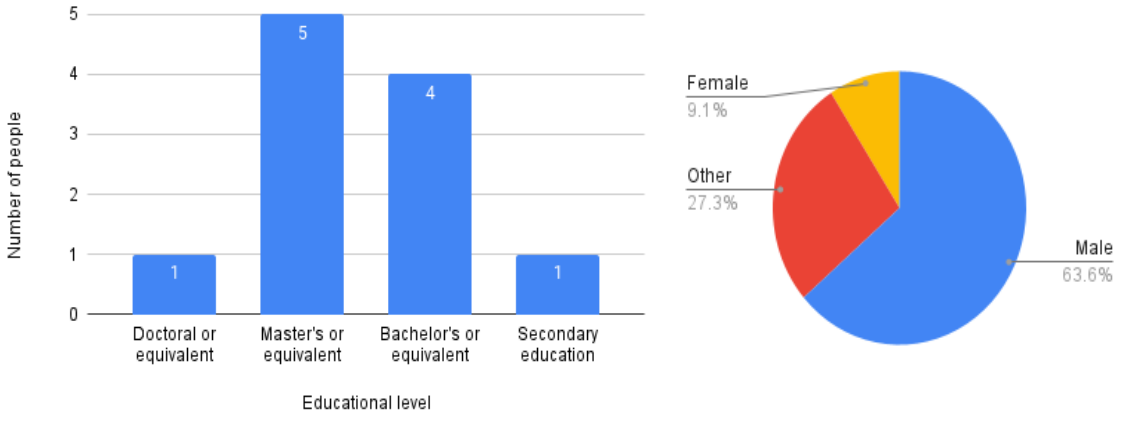

Figure 13: Educational level and gender distribution

Similar to the previous two languages, we present some of the sentences that the

participants were asked to rank (with their respective translation to English) and their corresponding scores. The full list of sentences is attached in the Appendix.

Example 7.
Original Sentence: "Die Top Ten der **meistgesuchten** Begriffe des Jahres 2004 sind demnach: 1."

Original Sentence (English): "The top ten most **searched** terms of 2004 are: 1."

Simplified Sentence: "Die Top Ten der **häufigsten** Begriffe des Jahres 2004 sind demnach: 1."

Simplified Sentence (English): "The top ten most **common** terms in 2004 are: 1."

Example 10.
Original Sentence: "Aufgrund des großen **Andrangs** der Nürnberger Bevölkerung fand jedoch von 11:00 Uhr bis 16:00 Uhr alle 15 Minuten eine Führung statt . "

Original Sentence (English): "Due to the large **number** of people in Nuremberg, however, a tour took place every 15 minutes from 11:00 am to 4:00 pm."

Simplified Sentence: "Aufgrund des großen **Interesses** der Nürnberger Bevölkerung fand jedoch von 11:00 Uhr bis 16:00 Uhr alle 15 Minuten eine Führung statt . "

Simplified Sentence (English): Due to the great **interest** of the Nuremberg population, however, a tour took place every 15 minutes from 11:00 a.m. to 4:00 p.m. "

Table 8: Evaluation results for the German Examples: 7. and 10. using the Likert scale [37] from 1 to 5

Parameter	Example 7.	Example 10.
Fluency	4.45	4.64
Meaning	3.82	4.18
Simplicity	4.27	4.45

The score for Fluency, Meaning and simplicity of all the sentences is summarized in the next graph:

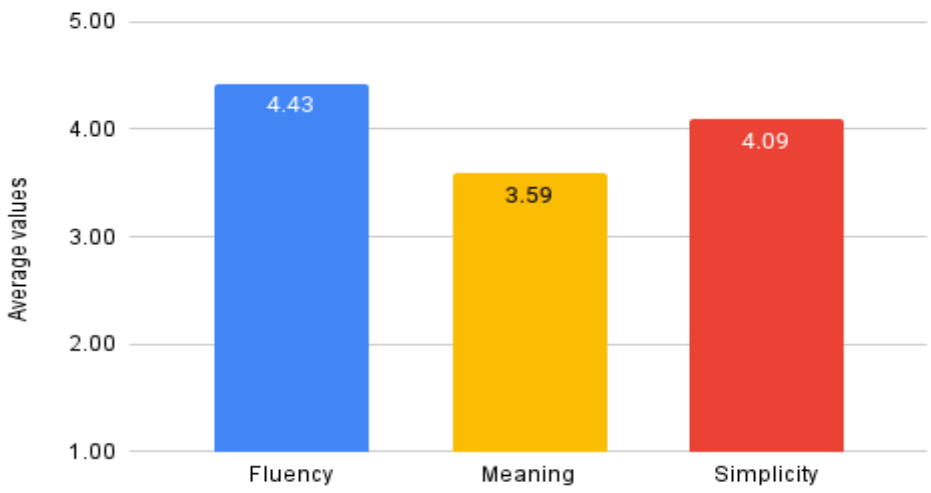

Figure 14: Average score of the 10 pairs of sentences for the values of Fluency, Meaning and Simplicity.

Overall we observe that for the German language the highest value corresponds also to the Fluency parameter, with a score of **4.42**. The Simplicity score is **4.09**. And finally the participants gave a score of **3.59** for how much of the original content was preserved during the sentence transformation.

Systems comparison

In the following table the metrics Fluency, Meaning and Simplicity are presented for the three languages using the same framework as in the individual reports:

Table 9: Average values of Fluency, Meaning and Simplicity for English, Spanish and German

Metric	English	Spanish	German
Fluency	4.08	3.81	4.43
Meaning	2.82	3.27	3.60
Simplicity	3.39	3.44	4.10

.

It is observed that for all the language models, the higher metric is Fluency. This metric commonly contrasted with language accuracy, moreover, it can be extended to include oral language proficiency. This metric alone cannot provide a complete overview of simplicity, however, a higher value is desired, as it is linked to linguistic knowledge and performance skills [38].

The Meaning metric, indicates how much of the original content is preserved after the word is processed by the system. A high value is desired, the contrary implies that the system replaced for a word that could be simple, but changes the general idea of the sentence, and affecting the global perception of simplicity.

Finally, the simplicity metric refers to multiple psycholinguistic features, such as **correctness**, which is the level of abstraction of the concept that the word attempts to describe, **Imageability**, or the capacity of the word to arouse mental images, **familiarity**, or level of exposure to the target word [39]. In this case the metric definition may vary for each of the participants, but overall, they will all perceive simplicity in a similar way.

Chapter 4

Discussion and Conclusions

4.1 Discussion

For the current work, we can point out that *word frequency*, has been shown to be correlated with the word's simplicity and with people's knowledge of that particular word [34]. However, this feature should not be used alone, as most of the times word frequency is calculated from a limited corpora, not the full language itself [15].

As it was shown in [39] the human perception of simplicity is influenced by several psycholinguistic features, one of them is concreteness, or the level of abstraction of the word for a specific event, implying that the amount of information that is added or removed to the word influences this metric. During the evaluation of the model, it was observed that even if some sentences obtained a high value of Fluency and Simplicity, the Meaning metric indicated that for that specific word, the best action was not to replace the word, so perhaps some additional constrains are required to prevent the replacement of the word when the meaning is degraded.

Including other stages, for instance, a *semantic similarity analysis* in the system may improve the meaning preservation of the simplified sentence, by comparing the cosine similarity of the original sentence with a series of candidate sentences, may improve its performance [10].

4.2 Conclusions

In conclusion, it would appear that the Lexical Simplification tasks greatly benefits from the word representation as a vector in the embedding space, by including them in the detection and the generation stages, however, in this stage the data supports the premise that the most challenging aspect is meaning preservation.

Our survey suggest that we still have a long way to go to balance simplicity and meaning. Despite the limitations, the system offers a mechanism that performs Lexical Simplification in Natural Language that is acceptable for the annotators. By using a Logistic Regression model, we found that it is possible to classify words in complex or not complex.

We explored the effects of including words embedding as a feature to the CWI system, and the results were statistically better for English and Spanish when compared using the features (word length and average word length). However, significant differences for the German language remained. This result leads to believe that other features may be more relevant for different languages, and further research is required to find the best features for each language.

Ideally, these findings could be replicated in a study where different algorithms are used to detect complex words, i.e. different machine learning systems. Furthermore, testing with different language models, and evaluate how they perform in different domains would cast new light on the language model selection.

We believe that apart from looking for which words should be replaced, future research should look to accept or not the change based on the meaning preservation.

List of Figures

List of Tables

Bibliography

[1] Saggion, H. *Automatic Text Simplification*. Synthesis Lectures on Human Language Technologies (Morgan & Claypool Publishers, 2017). URL `https://doi.org/10.2200/S00700ED1V01Y201602HLT032`.

[2] Drndarević, B., Štajner, S., Bott, S., Bautista, S. & Saggion, H. Automatic text simplification in spanish: A comparative evaluation of complementing modules. In Gelbukh, A. (ed.) *Computational Linguistics and Intelligent Text Processing*, 488–500 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013).

[3] Inui, K., Fujita, A., Takahashi, T., Iida, R. & Iwakura, T. Text simplification for reading assistance: a project note. In *Proceedings of the second international workshop on Paraphrasing*, 9–16 (2003).

[4] Alva-Manchego, F. E., Martin, L., Scarton, C. & Specia, L. EASSE: easier automatic sentence simplification evaluation. *CoRR* **abs/1908.04567** (2019). URL `http://arxiv.org/abs/1908.04567`. `1908.04567`.

[5] Leminen, M., Leminen, A., Smolander, S., Arkkila, E., Shtyrov, Y., Laasonen, M. & Kujala, T. Quick reorganization of memory traces for morphologically complex words in young children. *Neuropsychologia* (2020).

[6] Rello, L., Baeza-Yates, R., Bott, S. & Saggion, H. Simplify or help?: text simplification strategies for people with dyslexia. In Brajnik, G. & Salomoni, P. (eds.) *International Cross-Disciplinary Conference on Web Accessibility, W4A '13, Rio de Janeiro, Brazil, May 13-15, 2013*, 15:1–15:10 (ACM, 2013). URL `https://doi.org/10.1145/2461121.2461126`.

[7] Canning, Y., Tait, J., Archibald, J. & Crawley, R. Cohesive generation of syntactically simplified newspaper text. In Sojka, P., Kopeček, I. & Pala, K. (eds.) *Text, Speech and Dialogue*, 145–150 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2000).

[8] Saggion, H., Stajner, S., Bott, S., Mille, S., Rello, L. & Drndarevic, B. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Trans. Access. Comput.* **6**, 14:1–14:36 (2015). URL `https://doi.org/10.1145/2738046`.

[9] Paetzold, G. & Specia, L. Lexenstein: A framework for lexical simplification. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, 85–90 (2015).

[10] Bott, S., Rello, L., Drndarevic, B. & Saggion, H. Can spanish be simpler? lexsis: Lexical simplification for spanish. In Kay, M. & Boitet, C. (eds.) *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, 357–374 (Indian Institute of Technology Bombay, 2012). URL `https://aclanthology.org/C12-1023/`.

[11] Fellbaum, C. *WordNet*, 231–243 (Springer Netherlands, Dordrecht, 2010). URL `https://doi.org/10.1007/978-90-481-8847-5_10`.

[12] Bott, S., Rello, L., Drndarevic, B. & Saggion, H. Can spanish be simpler? lexsis: Lexical simplification for spanish. In *COLING* (2012).

[13] Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Stajner, S., Tack, A. & Zampieri, M. A report on the complex word identification shared task 2018. In Tetreault, J. R., Burstein, J., Kochmar, E., Leacock, C. & Yannakoudakis, H. (eds.) *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications@NAACL-HLT 2018, New Orleans, LA, USA, June 5, 2018*, 66–78 (Association for Computational Linguistics, 2018). URL `https://doi.org/10.18653/v1/w18-0507`.

[14] Yimam, S. M., Stajner, S., Riedl, M. & Biemann, C. Multilingual and cross-lingual complex word identification. In *RANLP*, 813–822 (2017).

[15] Horn, C., Manduca, C. & Kauchak, D. Learning a lexical simplifier using wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, 458–463 (The Association for Computer Linguistics, 2014). URL https://doi.org/10.3115/v1/p14-2075.

[16] Devlin, J., Chang, M., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR* **abs/1810.04805** (2018). URL http://arxiv.org/abs/1810.04805. 1810.04805.

[17] Wolf, T. *et al.* Transformers: State-of-the-art natural language processing. In Liu, Q. & Schlangen, D. (eds.) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, 38–45 (Association for Computational Linguistics, 2020). URL https://doi.org/10.18653/v1/2020.emnlp-demos.6.

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N. & Garnett, R. (eds.) *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008 (2017). URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[19] Face, H. Hugging-face: Fill-mask task pipeline. URL https://huggingface.co/models?pipeline_tag=fill-mask.

[20] Pires, T., Schlinger, E. & Garrette, D. How multilingual is multilingual bert? In Korhonen, A., Traum, D. R. & Màrquez, L. (eds.) *Proceed-*

ings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, 4996–5001 (Association for Computational Linguistics, 2019). URL `https://doi.org/10.18653/v1/p19-1493`.

[21] Wu, S. & Dredze, M. Are all languages created equal in multilingual bert? *CoRR* **abs/2005.09093** (2020). URL `https://arxiv.org/abs/2005.09093`. `2005.09093`.

[22] Rönnqvist, S., Kanerva, J., Salakoski, T. & Ginter, F. Is multilingual BERT fluent in language generation? In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, 29–36 (Linköping University Electronic Press, Turku, Finland, 2019). URL `https://aclanthology.org/W19-6204`.

[23] Abu-Mostafa, Y. S., Magdon-Ismail, M. & Lin, H.-T. *Learning from data*, vol. 4 (AMLBook New York, NY, USA:, 2012).

[24] Bollapragada, R., Nocedal, J., Mudigere, D., Shi, H.-J. & Tang, P. T. P. A progressive batching l-BFGS method for machine learning. In Dy, J. & Krause, A. (eds.) *Proceedings of the 35th International Conference on Machine Learning*, vol. 80 of *Proceedings of Machine Learning Research*, 620–629 (PMLR, 2018). URL `http://proceedings.mlr.press/v80/bollapragada18a.html`.

[25] Cortes, C., Mohri, M. & Rostamizadeh, A. L2 regularization for learning kernels. *CoRR* **abs/1205.2653** (2012). URL `http://arxiv.org/abs/1205.2653`. `1205.2653`.

[26] Piantadosi, S. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review* **21**, 1112–1130 (2014).

[27] Goldberg, Y. Neural network methods for natural language processing. *Synthesis lectures on human language technologies* **10**, 1–309 (2017).

[28] Qiang, J., Li, Y., Zhu, Y., Yuan, Y. & Wu, X. Lsbert: A simple framework for lexical simplification. *ArXiv* **abs/2006.14939** (2020).

[29] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A. & Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *arXiv preprint arXiv:1506.06724* (2015).

[30] Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H. & Pérez, J. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020* (2020).

[31] Cañete, J. Compilation of large spanish unannotated corpora (2019). URL `https://doi.org/10.5281/zenodo.3247731`.

[32] Chan, B., Schweter, S. & Möller, T. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, 6788–6796 (International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020). URL `https://aclanthology.org/2020.coling-main.598`.

[33] Ostendorff, M., Blume, T. & Ostendorff, S. Towards an open platform for legal information. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, 385–388 (Association for Computing Machinery, New York, NY, USA, 2020). URL `https://doi.org/10.1145/3383583.3398616`.

[34] Leroy, G., Endicott, J. E., Kauchak, D., Mouradi, O. & Just, M. User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning, and information retention. *Journal of Medical Internet Research* **15** (2013).

[35] Yacouby, R. & Axman, D. Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, 79–91 (Association for Computational Linguistics, Online, 2020). URL `https://aclanthology.org/2020.eval4nlp-1.9`.

[36] Alva-Manchego, F. E., Martin, L., Bordes, A., Scarton, C., Sagot, B. & Specia, L. ASSET: A dataset for tuning and evaluation of sentence simplifica-

tion models with multiple rewriting transformations. In Jurafsky, D., Chai, J., Schluter, N. & Tetreault, J. R. (eds.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 4668–4679 (Association for Computational Linguistics, 2020). URL `https://doi.org/10.18653/v1/2020.acl-main.424`.

[37] Joshi, A., Kale, S., Chandel, S. & Pal, D. K. Likert scale: Explored and explained. *British Journal of Applied Science & Technology* **7**, 396 (2015).

[38] Chambers, F. What do we mean by fluency? *System* **25** (1997).

[39] Biran, O., Brody, S. & Elhadad, N. Putting it simply: a context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 496–501 (2011).

Appendix A

First Appendix

A.1 Simplified sentences from the English dataset

Example 1.

Reference: "Aegyptosaurus was a close relative of Argentinosaurus, a muchlarger dinosaur found in South America. "

Second Sentence: "He was a close friend of Argentino, a much larger dinosaur found in South America ."

Example 2.

Reference: "A man and a woman questioned on suspicion of assisting an offender have been released. "

Second Sentence: "A man and a woman arrested on charge of killing an animal have been arrested"

Example 3.

Reference: "#5-11 His government stands accused by Human Rights Watch of not taking adequate measures to protect the nation's citizens."

Second Sentence: "#5-11 His government stands accused by Human Rights Watch of not taking enough measures to protect the American citizens."

Example 4.

Reference: "#The neighbor gave one of them a towel for his injuries shortly before their mother came outside and called the children home."

Second Sentence: "#The neighbor gave one of them a towel for his cuts shortly before their mother came outside and called the children home"

Example 5.

Reference: "A SWAT team entered the home and found Thomas dead from a self-inflicted gunshot wound and her three other children dead."

Second Sentence: "A SWAT team left the home and found Thomas dead from a similar gun shot and her three other children dead."

Example 6.

Reference: "Google, which will be the newest entrant to the handset market, announced plans for the acquisition last year in a bid to secure Motorola's valuable patents and pave the way for a pairing of Google's Android mobile software and Motorola's handset business."

Second Sentence: "Google, which will be the next entry to the Android market made plans for the takeover last year in a bid to secure the key technologies and open the way for a partnership of the Android mobile software and the hardware business."

Example 7.

Reference: "It's a complete tragedy for the town."

Second Sentence: "It's a complete loss for the town."

Example 8

Reference: "A reform package for Spain's troubled banks, aimed at convincing investors that the sector is solvent and the country has a strategy to avoid a bailout, is expected to be released."

Second Sentence: "A new plan for the failed banks, aim at telling investor that the sector is liquid and the country has a plan to avoid a crisis is expected to be published"

Example 9.

Reference: "Alhamadee, who is from a village near al-Tamana, said sectarian tensions were low before the uprising, but have deteriorated as Sunni villages like al-Tamana joined the anti-Assad uprising."

Second Sentence: "Ali who is from a village near al-Tamana, said the relations were low before the revolution but have developed as Sunni villages like al-Tamana joined the popular movement"

Example 10.

Reference: "Authorities quickly moved in as emergency service breached the back door"

Second Sentence: "People quickly moved in as the service reached the back door"

Appendix B

Second Appendix

B.1 Simplified sentences from the Spanish dataset

Example 1

Reference: "Thomas Michael Fletcher (17 de julio de 1985, Harrow, Londres), conocido comúnmente como Tom Fletcher, es un músico y compositor británico cofundador de la banda británica McFly, completada por Danny Jones, Dougie Poynter y Harry Judd."

Second Sentence: "Thomas Michael Fletcher (17 de julio de 1985, Harrow, Londres), conocido comúnmente como Tom Fletcher, es un músico y compositor británico fundador de la banda británica McFly, completada por Danny Jones, Dougie Poynter y Harry Judd."

Example 2

Reference: "Mokujin es un muñeco hecho de madera de roble de hace 2000 años, de los bosques antiguos de Japón, donde se especulaba la presencia de algo mágico."

Second Sentence: "Mokujin es un muñeco hecho de madera de roble de hace 2000 años, de los bosques antiguos de Japón, donde se creía la presencia de algo mágico."

Example 3

Reference: "Hay abundantes pastos, fundamentalmente de maíz y cereales de secano, así como encinas."

Second Sentence: "Hay abundantes pastos, principalmente de maíz y cereales de invierno así como encinas."

Example 4

Reference: "Durante todo el siglo XVIII fue la única parroquia del municipio de Taüll, mientras que la de San Clemente ejercía de capilla del cementerio moderno de Taüll, pero en los siglos anteriores, habían compartido parroquialidad las dos iglesias."

Second Sentence: "Durante todo el siglo XVIII fue la única parroquia del municipio de Taüll, mientras que la de San Clemente ejercía de capilla del cementerio moderno de Taüll, pero en los siglos anteriores habían compartido parroquia las dos iglesias."

Example 5

Reference: "Por último,tras disfrutar del primer y más corto pasacalles del Carnaval, se procede a rifar de manera gratuita como muestra de agradecimiento entre los asistentes de la "Cesta del Carnaval", que incluye un kit de supervivencia para que al premiado/da no le falte de nada para poder pasar estas fechas."

Second Sentence: "Por último,tras disfrutar del primer y más corto desfile del Carnaval, se procede a comer de manera gratuita como muestra de agradecimiento entre los asistentes de la "Cesta del Carnaval", que incluye un kit de supervivencia para que al ganador no le fal de nada para poder pasar estas fechas."

Example 6

Reference: "El Heinz Field es un estadio de fútbol americano situado en la ciudad de Pittsburgh (Pensilvania), Estados Unidos."

Second Sentence: "El Heinz Field es un estadio de fútbol americano situado en la ciudad de Pittsburgh , Estados Unidos."

Example 7

Reference: "Fue nominada en dos oportunidades como mejor actriz para los premios Martín Fierro, por su labor como mejor actriz en las telenovelas "Amor en custodia" y "Sos mi vida"; y como revelación en cine para el Premio Cóndor de Plata por la película de Sergio Renán, "Tres de corazones", en la pantalla grande."
Second Sentence: "Fue nominada en dos oportunidades como mejor actriz para los premios Martín Fierro, por su labor como mejor actriz en las películas "Amor en custodia" y "Sos mi vida"; y como revelación en cine para el Premio Cóndor de Plata por la película de Sergio Renán, "Tres de corazones", en la pantalla grande."

Example 8

Reference: "Las dos fueron erigidas y consagradas al mismo tiempo: la de Santa María fue consagrada un día más tarde que la de San Clemente, en el año 1123."
Second Sentence: "Las dos fueron erigidas y dedicadas al mismo tiempo: la de Santa María fue consagrada un día más tarde que la de San Clemente, en el año 1123."

Example 9

Reference: "Cuando se ha conseguido la textura apropiada se envasa y se distribuye como un producto listo para el consumo."
Second Sentence: "Cuando se ha conseguido la textura apropiada se vende y se distribuye como un producto listo para el consumo."

Example 10

Reference: "Es una persona sarcástica que gusta de hacer bromas pesadas, pero puede engañar fácilmente a los demás, tiene esa cara que refleja también engaño e hipocresía, por lo que jamás sabrás que es lo que está pensando en realidad."
Second Sentence: "Es una persona divertida que gusta de hacer bromas pesadas, pero puede engañar fácilmente a los demás, tiene esa cara que refleja también engaño e ignorancia por lo que jamás sabrás que es lo que está pensando en realidad."

Appendix C

Third Appendix

C.1 Simplified sentences from the German dataset

Example 1

Reference: "Roland Freyer behauptet in seiner Anzeige , die Ausstellung und damit die insgesamt 160 Kopien der Terrakotta-Tonkrieger würden ohne chinesisches Einverständnis gezeigt ."

Second Sentence: "Roland Freyer behauptet in seiner Anzeige , die Ausstellung und damit die insgesamt 160 Kopien der Ausstellung würden ohne Chinas Einverständnis gezeigt ."

Example 2

Reference: "Zu den weiteren Bereichen , die ausgewertet wurden , zählen zum Beispiel auch die populärsten Gerichte , die meistgesuchten Elektronikgeräte und auch die Top Ten der TV-Shows . "

Second Sentence: "Zu den weiteren Bereichen , die untersucht wurden , zählen zum Beispiel auch die bekanntesten Gerichte , die bekanntesten Geräte und auch die Top Ten der TV-Shows . "

Example 3

Reference: "Meyer hatte , obwohl er immer sehr aggressiv gegen diverse Affären von Sozialdemokraten vorging , wärend seiner Arbeit seit Oktober 2000 als CDU-Generalsekretär und engster Mitarbeiter von Angela Merkel immer wieder Zahlungen von der RWE bekommen ."

Second Sentence: "Meyer hatte , obwohl er immer sehr aktiv gegen diverse Affäre von SPD vorging , wärend seiner Arbeit seit Oktober 2000 als Generalsekretär und engster Mitarbeiter von Angela Merkel immer wieder Zahlungen von der RWE bekommen ."

Example 4

Reference: "Wer an den Kursen nicht teilnehme , soll den Anspruch auf seine Aufenthalts- und Arbeitsgenehmigung verlieren ."

Second Sentence: "Wer an den Kursen nicht teilnehme , soll den Anspruch auf seine Einreise und Genehmigung verlieren ."

Example 5

Reference: "Dezember der Regierung Janukowitschs das Misstrauen ausgesprochen , jedoch ist dessen Votum rechtlich nicht bindend und (Noch-) Präsident Kutschma hat sich bisher geweigert diesem Votum zu entsprechen und die Regierung zu entlassen . "

Second Sentence: "Dezember der Regierung Janukowitschs das Misstrauen ausgesprochen , jedoch ist dessen Urteil rechtlich nicht bindend und der (Noch-) Präsident s hat sich bisher geweigert diesem Urteil zu entsprechen und die Regierung zu entlassen . "

Example 6

Reference: "September 2004 festgestellt worden , als Wiktor Juschtschenko ins Krankenhaus eingeliefert wurde ."

Second Sentence: "September 2004 festgestellt worden , als Wiktor Juschtschenko ins Krankenhaus gebracht wurde ."

Example 7

Reference: "Die Top Ten der meistgesuchten Begriffe des Jahres 2004 sind demnach: 1."

Second Sentence: "Die Top Ten der häufigsten Begriffe des Jahres 2004 sind demnach: 1."

Example 8

Reference: Es geht darin um Themen aus Geschichte , Literatur , Philosophie , Kunst und Musik , die nach Meinung von Schwanitz zur Allgemeinbildung in Deutschland gehören sollten ."

Second Sentence: "Es geht darin um Themen aus Geschichte , Literatur , Philosophie , Kunst und Musik , die nach Meinung von Shwanitz zur Bildung in Deutschland gehören sollten ."

Example 9

Reference: "Eine Schwächung des Präsidentenamtes wird jedoch von der Opposition mit ihrem Präsidentschaftskandidaten Juschtschenko strikt abgelehnt ."

Second Sentence: "Eine Besetzung des Amtes wird jedoch von der Opposition mit ihrem Kandidaten Juschtschenko strikt abgelehnt ."

Example 10

Original Sentence: "Aufgrund des großen Andrangs der Nürnberger Bevölkerung fand jedoch von 11:00 Uhr bis 16:00 Uhr alle 15 Minuten eine Führung statt . "

Second Sentence: Aufgrund des großen Interesses der Nürnberger Bevölkerung fand jedoch von 11:00 Uhr bis 16:00 Uhr alle 15 Minuten eine Führung statt . "