Master's Thesis

Generating English explanations of logical formulas: measuring the quality of the generated sentences

Igor Kuczuk Modenezi

First Supervisor:

Prof. dr. C.J. (Kees) van Deemter

Second Supervisor:

Dr. G.A.W. (Gerard) Vreeswijk

Student Number 6889689

MSc in Artificial Intelligence

Utrecht University

April 2022

Abstract

Logic languages such as First Order Logic can describe complex ideas, with the downside of being too abstract. In order to make them clearer, it is possible to translate sentences in them to Natural Language. In order to examine this kind of system further, we chose to focus on one domain: Tarski's World. We created a system that has a First Order Logic formula that has a quantifier as input and an English sentence as output. Our focus was generating a more natural-sounding sentence and explore the quality of the sentence generated. This was possible through the development of two metrics: Naturalness (how natural a sentence is) and Clarity (how grammatically clear a sentence is). Both showed promising results but were ultimately flawed and required further improvements. During the development of the metrics, it became apparent that this type of sentence construction presents scope ambiguity. In order to determine the acceptable readings, a survey was conducted. From the results we concluded a few things: that sentences with adjectives had nocuous ambiguity; quantifiers do not present scope ambiguity; and sentences that contain numbers instead of quantifiers present scope ambiguity.

Table of Contents

1. Introduction 5
2. Related Work 6
   2.1. Generation From Logic 6
      2.1.1. Tarski's World 6
      2.1.2. Hatzilygeroudis and Mpagouli (2007) 6
      2.1.3. Flickinger (2016) 7
      2.1.4. Conclusions 8
   2.2. Quality of the Generated Output 8
      2.2.1. Mayn and van Deemter (2020) 8
      2.2.2. Khan and van Deemter (2012) 9
      2.2.3. Conclusions 10
3. Initial Algorithm 10
   3.1. Formula Fragment 10
   3.2. Tree Construction 11
   3.3. Sentence Generation 11
   3.4. Background Knowledge Substitution 12
   3.5. Conclusions and Future Work 12
4. Metrics 12
   4.1. Naturalness 13
      4.1.1. Perplexity 13
      4.1.2. SLOR 14
      4.1.3. Readability Metrics 14
      4.1.4. Conclusions 15
   4.2. Clarity 15
      4.2.1. Conclusions 17
5. Experiment: Perceived ambiguity of scopally ambiguous sentences 17
   5.1. Questions 18
   5.2. Design and Materials 18
   5.3. Procedure and Participants 20
   5.4. Results 21
      5.4.1. "All red squares and circles are on the left" 22
      5.4.2. "All squares and circles that are red are on the left" 22
      5.4.3. "All red squares or circles are on the left" 23
      5.4.4. "All squares or circles that are red are on the left" 23

5.4.5. “All squares and circles are on the left” 23
5.4.6. “All squares or circles are on the left” 24
5.4.7. “X squares and circles are on the left” 25
5.4.8. “X squares or circles are on the left” 27
5.5. Conclusions from the experiment 29
6. Conclusions 29
7. References 31

1. Introduction

When first learning Logic Languages, a student can comprehend easily how small fragments can be translated but usually struggles when the composition is larger and the relations of the parts are more complex. Said step is similar to learning a foreign language. Unlike a foreign language, the syntax of Logic Languages is not immediately apparent. Seeking to bridge the gap between the abstractness of logic to a clear translation between languages, a common practice is to translate back to Natural Language using a series of simplifications and analogies to the fragments (such as presented in Hatzilygeroudis and Mpagouli (2007)). Since the normal procedure requires manual translation, an automated system would be extremely useful.

An initial idea might be use to Propositional Logic, since it is good to describe simple and concise ideas. The problem is that it is limited by not being able to use quantifiers, limiting how expressive and precise it is. It is reasonable, then, that the Logic Language used is one that presents quantifiers. According to Barwise, J., & Etchemendy, J. (1993), First Order Logic has the same foundations as Propositional Logic but includes this extra layer of depth. Therefore, an effort was made to develop automated systems to translate First Order Logic (FOL) into Natural Language.

This is not as sought after as Propositional Logic because it creates a complexity that can be difficult to deal with when the sentences became too long. For this reason, it is important to limit the scope of the project to a specific context. For the current project, it was decided to work with Tarski's World (first presented in Barwise, J., & Etchemendy, J. (1993)), a domain that has limited use but allows for a good first step for a broader system. It should be pointed out that there is very limited related work using FOL and this domain (Karttunen (1989), Flickinger (2016)).

The objective of the current work is to explore this gap currently present in the field, by developing logic to language generation system that uses quantifiers in the formula, with naturalistic sentence construction. Through the development of this, it was possible to observe specific structures that would lead to scope ambiguity and different possible readings. For this reason, it is important to explore and determine whether the average reader can infer the best meaning even with the presence of some ambiguity.

Since there is some literature referring to translating logic to Natural Language, an existing system was used as the basis. This way, with a good, working foundation it is possible to obtain a functional system with less trial and error.

It should also be pointed out that it was programmed using easy-to-access libraries, making it easy and accessible for future work. The programming language used was Python, since it is easy to work with and has an ample variety of resources available to work with.

In the following sections, further details of the project itself and its specific context will be discussed. In Section 2, we approach similar works and other resources necessary for this project. Section 3 describes the algorithm and the technical details of the program. In Section 4, we deal with a few metrics tested and further work necessary to generate a good sentence. Section 5 presents a survey developed to determine the possible interpretations of a single sentence and whether a certain amount of ambiguity is acceptable without losing meaning. Finally, Section 6 presents the overall conclusions and recommendations for further development.

2. Related Work

2.1. Generation From Logic

In order to develop a new system, it is necessary to go through some of the previous efforts of generating language from Logic. This way, it is possible to observe what works and what needs to be expanded upon.

2.1.1. Tarski's World

Considering logic and language, it is almost infinite the number of possible sentences. Therefore, it is more practical to delimit to a specific type of world in which to make the analysis. Said world is called a "domain" in logic.

In 1993, Jon Barwise and John Etchemendy presented a domain called Tarski's World, to teach First Order Logic at Stanford University. It is comprised of three-dimensional objects situated in a plane. This limits the vocabulary to adjectives that can be given to objects (such as size and color), type of figures (whether it is a cube, tetrahedron, and so on), positional information (if something is to the left or right; if it is nearby or far, and so on) and several objects (including quantifiers). With this delimitation, it is possible to create formulas that are simple enough that a student can learn First Order Logic, but complex enough to test logic systems. An example of the system can be seen in Figure 1.

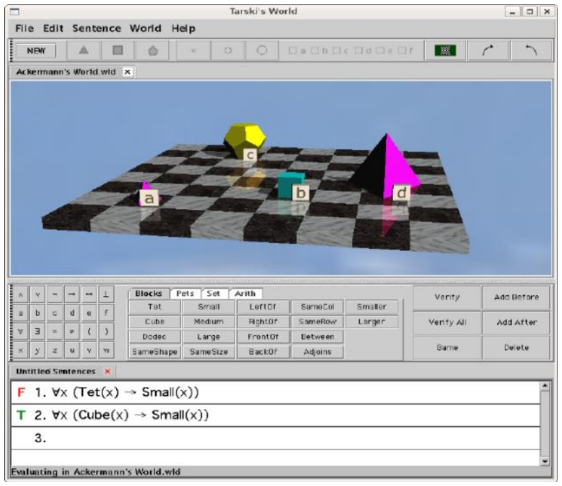

Figure 1 – An example of Tarski's World being used

It is possible to observe that such a domain is a good starting point for a project of translating First Order Language to Natural Language for multiple reasons: it was designed with this logic in mind; it is expressive enough to form complex formulas but with a vocabulary that is limited enough so it does not get unfeasible; and it has a direct visual analog, permitting an easy way of presenting a formula (or clarifying ambiguity).

2.1.2. Hatzilygeroudis and Mpagouli (2007)

To assist classes of Artificial Intelligence at the University of Patras, a web-based intelligent tutoring system was developed. It was observed that the students struggled to translate from Natural Language to FOL. It was decided, then, that the system could translate the FOL formula to Natural Language, for the students to determine if their initial formula fits with the desired translation and make further alterations before submitting it.

According to the Related Work section, there was no similar literature depicting this effort of translating from FOL to Natural Language. It should be pointed out, though, that a few other related works were mentioned, where small parts were used as inspiration to develop their algorithm.

Named FOLtoNL conversion algorithm, the system was designed with the following formula in mind:

(Quantifiers) (Predicates) => (Predicates)

Where "=>" represents implication, "Quantifiers" can use any quantifier (negated or not), and "Predicates" can be predicates using connectives. Since the system is not interested in how natural the phrasing is, the predicates and quantifiers are presented as plainly as possible. So, for example, if the input is: (forall x) (Squar x) => (Red x) . The output would be: "For all x, if x is square then x is red". Considering the objective of the project, it is enough to achieve it and a good first step in solving the problem. Some other examples can be seen in Table 1.

Input	Output
(forall x)(forall y) likes(x,y) => appeals(y,x)	**forall x and y if x likes y then y appeals to x**
(forall x) bat(x) => ~feathered(x)	**forall x if x is a bat then x is not feathered**
(forall x)(loves(father(x),x)&loves(mother(x),x))	**forall x the father of x loves x and the mother of x loves x**
~(forall x)(exists y)cares(x,y)	**for some but not all x , exists y such that y cares about x**
is(sum(2,3),5)	**the sum of 2 and 3 is 5**
(forall x)(exists y)(exists z)(exists w) human(x) => name(y,x) & age(z,x) & birthday(w,x)	**forall x exist y , z and w such that if x is a human then y is the name of x and z is the age of x and w is the birthday of x**
(forall x) bird(x) & ~flies(x) &swims(x) => penguin(x)	**forall x if x is a bird and x does not fly and x does swim then x is a penguin**

Table 1 – Examples of sentences generated by FOLtoNL. It interprets complex formulas, with the drawback of not presenting them in a naturalistic way

Given the way the predicates are presented in this system, the authors impose the limitation that each "Predicates" fragment could only use one type of connective. This was made to avoid any type of ambiguity, with the cost of expressivity.

From this work, several things can be concluded: it is possible to make a FOL to Natural Language translation; using implication in the formula can be very expressive; and some form of method to include multiple connectives without becoming too ambiguous can be beneficial for the system.

2.1.3. Flickinger (2016)

Instructional software, such as Tarski's World, are useful tools for teaching logic (and FOL). However, the way sentences are presented in NL can be limited using them. For this reason, the author proposes an English generator that produces a series of paraphrases, covering the possible representations and meanings desired.

The input of the system is a well-formed FOL expression, without quantifiers. The said decision is not discussed in the paper, but it can be concluded that it is because of the added complexity generated by quantifiers.

The algorithm works in the following manner: the input is converted to a grammar-specific semantic representation, followed by the generation of the paraphrases. From the FOL expression, it is converted to a structure that uses the Minimal Recursion Semantics framework, a previously developed framework within the English Resource Grammar (one of the resources used by the author). Simply, this process turns each part of the expression into a separate object, with the associated word, argument, and grammar function.

From the said framework, it is possible to already generate sentences, albeit trivial ones. To expand its function, it was necessary to develop an MRS-to-MRS mapping, enabling the creation of multiple paraphrases. For this, a series of rules are necessary. Through an analysis of the textbook that accompanies Tarski's World, it was possible to determine a set of 143 paraphrase rules.

Said the high number of variations is necessary for simplification and formulating more complex formulas. The simplifications are essential to eliminate undesirable readings from the original expression, given the difference between the representation in logic to the representation in Natural Language.

Regarding the limitations of the system, the sentence structure is generally of the type "a is large and b is large" (presence of variables), with possible simplifications. Given that the system's intent is to be used for the assistance of students, it is acceptable for the presentation to have variables, although not common when being used in regular conversation. It should also be noted that, since it does not use quantifiers, the system is limited to a subset of possible expressions that can be expressed in the domain.

2.1.4. Conclusions

From the previous work, it was possible to observe what are the gaps in the field. Hatzilygeroudis and Mpagouli (2007) provided a appropriate system, but with no quantifiers, limiting the possible sentences constructed. Flickinger (2016) also avoided the same thing, pointing to a gap in the literature for this type of sentence. It was also observed that, for both of the previously mentioned works, the sentences have the presence of variables ("x", "y" or "z", as presented in their examples), creating sentences that are functional but do not sound natural.

For this reason, the current project aims at addressing both of these limitations, therefore making a valuable expansion on the current body of work.

2.2. Quality of the Generated Output

Simply generating a sentence is not enough, since it could be unintelligible or confusing. For this reason, it is necessary to explore a few ways that some other work studied how to generate a good output. From this, it will be possible to have some ideas of what to do and what to expand upon.

2.2.1. Mayn and van Deemter (2020)

In the paper "Towards Generating Effective Explanations of Logical Formulas: Challenges and Strategies", the authors focus on the generation of sentences that are helpful for the user and similar to phrases used regularly in English. The domain used was Tarski's World.

To verify the challenges and the procedures in generating sentences, a system was created, where the input is a propositional logic formula and the output is a natural language sentence. It functions in the following way: the formula is parsed into a tree structure; if possible, the tree is simplified using a set of rules; it then constructs the output by reading the tree from left to right.

Throughout the development and testing of the program, a few questions were raised regarding the quality of the sentence generated: a single sentence could be ambiguous, with a preferred reading but with multiple ways of writing it; and the readability of a sentence, which can be unambiguous but sound unnatural. These aspects should be considered for the further development of the system.

According to Swets et al. (2008), ambiguous sentences can be processed faster than unambiguous ones, when no in-depth information is needed. For this reason, the paper addresses aggregation and how it affects the readability of a sentence. This means that, instead of expressing each characteristic of an object as a

separate sentence, for example, it is possible to present it as a single sentence with a complex sentence structure. An example of this can be seen below, where the second sentence has aggregation:

x is not a cube and it is not the case that x is smaller than y or x is the same shape as y.

x is not a cube, nor is it smaller than or the same shape than y.

It should also be noted that the sentence is clearer. This points to the fact that a longer sentence does not mean clearer information, depending more on the case and sentence structure.

2.2.2. Khan and van Deemter (2012)

The article explores the effect of surface ambiguity in the comprehension of a sentence, from the perspective of language generation. To further clarify what that means, it is necessary to define some aspects of Language Generation.

From the design point of view, the generation of sentences has two important aspects: Content Determination and Linguistic Realization. The first can be simply defined as "what to say" and the second as "how to say". Thus, even if the content is well determined and is unambiguous, the form in which is expressed can generate ambiguity on its own. For this reason, it is important to mitigate it (and also determine what is a perceived ambiguity to a reader).

The article focuses on coordination ambiguities (presented in sentences of the form "the ADJ NOUN1 and NOUN2"), specifically scopally ambiguous sentences. This means that ADJ can have more than one scope: wide scope (where it can be applied to both nouns); or narrow scope (where it can only be applied to Noun1).

The main objective is to avoid ambiguities that cause misunderstanding. It is first necessary to determine which ones have that effect. The approach decided on was conducting three investigative studies: possible interpretations of noun phrases by readers; preference of readers regarding sentence length; and measuring the reading and comprehension times for sentences. A summary of each one will be presented below.

For the first study, the main aspect for analysis was which type of structures usually garner the interpretation of wide and narrow scope. This is determined by the strength of the influence of ADJ or Noun2 on Noun1. So, for example, if the influence of the adjective is stronger than from the other noun, the preferred reading would be narrow scope: (ADJ Noun1) and Noun2; wide scope would be when this influence is lower. Their findings were that the strength of the adjective is the deciding factor in which interpretation is preferred.

For the second study, the focus was on brevity and clarity. More specifically, if one is more important than the other. The results point out that brief sentences are preferred when they are clear and, given a situation where it is necessary to choose only one, the readers preferred clarity. This is relevant for Language Generation systems since it points out that readers do not have problems with a longer sentence as long as it has a clear meaning.

The final study is centered around reading and comprehension time, looking to determine the inherent preferences for how long and clear a sentence is. It was observed that there is a preference for brief sentences and also sentences with predictable structures. It should be pointed out that having both happening in a single sentence is possible but not guaranteed. The authors did not reach a conclusion about how to balance both, but reflected on it, mentioning that there might not be a definitive answer to this problem.

2.2.3. Conclusions

From Mayn and Deemter (2020), it was possible to observe that the tree structure used by the system is efficient and, therefore, will be used as the basis for our system. It was also questioned whether ambiguity has a strong effect on the comprehension of the sentence; and the importance of naturalness of the sentence construction. Since these are important to determine the quality of a sentence generated, they will be further explored by the current work.

From Khan and Deemter (2012), the topic of ambiguity was addressed again, with the focus on scopally ambiguous sentences, a construction that can happen in Tarski's World and in the way our system was devised. The conclusion was that the preferred meaning can be obtained from the strength of the adjective related. It was also explored the topics of brevity and clarity. They are relevant for generating good output and will be addressed in the current work.

From this set of works, it is possible to observe that ambiguity is also a part of the discussion surrounding the quality of a generated sentence. For this reason, it is necessary to explore how it affects the understanding of a given sentence and if it is necessary to avoid it.

3. Initial Algorithm

Given the similar works presented in Section 2, it becomes apparent the challenges and needs to be required for creating a FOL to Natural Language translation system. The system presented in Section 2.2.1 was used as the foundation for the current project.

Therefore, the first step was to adapt it from Propositional Logic to First-Order Logic. Given that both projects use the same domain, the predicates were maintained, as well as the connectives. This means that the main change was the inclusion of quantifiers, done in the formula parser.

Since the main objective of this work is to explore the quality of generated sentences in the given domain, it is necessary to develop an initial language generation system. This way, it is possible to test quality measures and sentence constructions that were created by a concrete algorithm. This also means that the issues mentioned before (clarity, naturalness, and ambiguity) will not be dealt with in this chapter, focusing on the algorithm alone.

3.1. Formula Fragment

To determine the scope of the project, we agreed on a formula fragment, on which to develop the system. Anything beyond it would be considered for further projects. Considering the major difference between Propositional Logic and FOL, the presence of a quantifier is considered essential. It was also observed that, when trying to describe objects and their characteristics, we commonly use implication, making it a central aspect of the fragment. The fragment can be seen below:

$$(Quantifier): (Predicates) \rightarrow (Predicates)$$

The "Predicates" statement refers to any number of predicates and connectives. The "Quantifier" statement indicates the use of one specific quantifier and using multiple ones should be the next step of a future expansion of the project.

To make it easier to parse, the sentence is written is with brackets, indicating the hierarchy of what was written. For example, if a part of the formula desired to express is: $Red(x) \wedge Ball(x) \vee Cube(x)$, it is necessary to define the scope of the connectives when writing the input. In this case, it could be: $Red(x) \wedge (Ball(x) \vee Cube(x))$. Therefore, the user is responsible for making it clear what type of sentence structure

is desired. Such a design decision allows the construction of a sentence that sounds more natural, with the setback of making the user have a clear understanding of what they want to express.

One necessary limitation of the fragment was the use of negation. Considering how it can change the structure of a sentence, it was determined that the negation can only affect one predicate. The application of negation with a wider scope is a big challenge on its own and will be developed in future work.

3.2. Tree Construction

After the formula is declared by the user, a tree is generated with the parser. From the root of the tree, the main branch is composed of the quantifier and the implication. Said decision was taken because they are the central aspects of the fragment and will appear in every input of the system, therefore it should be the central part of the tree. An example of a tree can be seen in Figure 2.

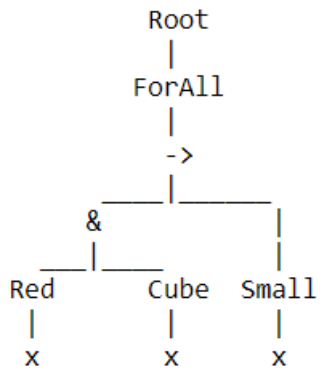

Figure 2 – Example of a parsed tree. The original formula is: ForAll: (Red(x) & Cube(x)) -> Small(x).

3.3. Sentence Generation

The way the sentence is generated is by using the main branch as the generic structure and building upon it. This way, all sentences generated have the general structure as follows: "(All/some) ____ are _____", where the spaces are filled with the remainder of the tree. Considering the example of Figure 2, the generated sentence would be "All red cubes are small". This specific wording was decided based on a common way of expressing things in English since the structure "if ___ then ___", although more commonly associated with implication, is also less fluid in a single sentence.

It should also be noted that the connectives can be suppressed when it is not needed. For example, "Red(x) & Cube(x)" is more naturally translated to "red cube" instead of "red and cube", therefore that is the preferred way of generating it. Said simplification is only used when there is no loss of meaning, thus it only happens with the conjunction connective but not with disjunction.

To implement this type of simplification, it was necessary to create attributes for the predicates. This means that predicates like "Cube" and "Tetrahedron" are defined as objects, "Red" and "Blue" as colors, and "Small" and "Medium" as size, for example. This way, it is possible to structure the sentence depending on what predicates are present, to make it sound naturalistic. If, for example, a predicate of type color or size appears without a predicate of type object, the word "objects" can be added (depending on the position in the sentence). This allows the creation of a series of rules, without having to manually specify every predicate.

3.4. Background Knowledge Substitution

While thinking about possible simplifications, we came up with the idea of using background knowledge to substitute negated predicates. What this means is that using the knowledge of the domain and the types of predicates, it is possible to substitute a negated predicate with a simple predicate. For example:

$$\neg\, Small(x) \equiv Medium(x)$$

So, instead of stating that something is not small, it can be stated that something is medium. The same process can be extended to types of objects and colors, as long as it has a counterpoint.

This is an interesting simplification since it does not require restructuring the sentence, just a substitution. Its use was also considered optional, in case it is not desired by the user.

Considering the limitations, the Background Knowledge Substitution can currently only be used with predicates that have a direct opposite. Using the example provided above, it is only possible to substitute "NOT Small" for "Medium" if there is no predicate "Big" included. This is because cases that have multiple possible alternatives would not generate a simplification, since it would require listing all of them.

3.5. Conclusions and Future Work

The current progress is a good first step into solving the problem of translating from FOL to Natural Language for Tarski's World domain. For future work, it is advised to expand to other domains and implement global negation.

Regarding the current work, the developed algorithm can be used to explore the previously mentioned quality concerns: Clarity, Naturalness, and ambiguity. For the next chapter, the main focus will be on the first two, given that they can be identified more easily, by using metrics, whereas ambiguity can be identified from the perception of the reader.

4. Metrics

When developing a language generation algorithm, it is important to evaluate the quality of the output. Creating a system that can generate English sentences can be done quickly, but if the output is confusing or not clear enough, it will not be useful. For this reason, it is important to explore quality metrics. As mentioned in Section 2.2, there are a few characteristics to account for. Even though all are important, exploring them all is unfeasible for the current scope. It should also be pointed out that most works related to language generation do not go too in-depth on this topic, therefore making exploration more relevant.

Considering the topics mentioned in Section 2.2, our exploration of the metrics is about clarity (proposed by Khan and Deemter (2012)) and naturalness (proposed by Mayn and Deemter (2020)). For this reason, two metrics were created and tested.

One of them is Naturalness. The idea is to employ a bigram Language Model to determine if the sequence of words is likely to happen. Since using the direct probability is affected by how long a sentence is, our solution was to test a few derived formulas that mitigate its effect of it and compare the results. This metric is important, given that it determines if a sentence sounds natural, making it easier for an average English reader to comprehend its intended meaning.

The other is Clarity. The basic idea is that a Syntactic Parser analyses the sentence (using grammar for the Tarski's World domain) and creates several possible syntactic structures, with the relative probability of each one. If a sentence has multiple structures, it should be more ambiguous than one with less. Although

similar to finding possible ambiguities, this metric is necessary to determine how many possible ways are to interpret a given sentence and, in doing so, find out if a sentence structure is grammatically clear.

4.1. Naturalness

The first step of this metric was to determine the length of the n-gram Language Model. Considering that a short sentence with our fragment can have four words (for example, "all cubes are red"), it should be between a trigram and a bigram. It was decided to use a bigram given that it does not limit as much the possible sentence structures.

To train the bigram, it was used a free online database available online that compiles a series of news articles. Since the vocabulary used in Tarski's World is not very usual, the database created for Clarity was included, to guarantee the appearance of all the words used (for example, "tetrahedron" is very uncommon and "tetrahedra" even more so). The training was done using the NLTK.lm library, with 43470198 ngrams obtained.

Using the ngram probabilities directly is not desired since sentences that are longer would present a lower score overall. For this reason, a series of existing scores were tested for our objective.

4.1.1. Perplexity

The first one tested was Perplexity. Used frequently to measure the quality of a Language Model, this metric is indirectly inversely proportional to the probability of a sentence occurring. For example, if a sentence has a low probability of occurring, it has high perplexity. The formula to obtain it can be written as follows:

$$Perplexity = \sqrt[Length]{\frac{1}{Probability\ (bigram)}}$$

Given that using root in programming can lead to rounding, it was decided to use the formula in a different format, defined as the inverse of cross-entropy. The formula is as follows:

$$Perplexity = 10^{-\frac{\log Probability(bigram)}{Length}}$$

Considering that it is related to the probability, it is expected to be slightly affected by the length of the sentence, even though it is divided by it.

When testing with a few sentences, the results were promising. "All cubes are red" got a score of 46.3, "All small cubes are red" got 47.2, "All cubes are red and all tetrahedra are red" got 53.2, and "All red cubes and tetrahedra are large" got 67.5. These initial results show that the score is affected by the length of the sentence, but that an ambiguous sentence got a significantly higher score.

To test if an unnatural sentence (or incomprehensible) gets a worse score, a few sentences with 7 words (as to compare to the previously tested ambiguous sentence) were tested. "All large tetrahedra and cubes are cubes" got a score of 73.7 and "All cubes tetrahedra and cubes are cubes" got 78.8. From this, it is possible to observe that sentences with the same length but worse constructions got worse scores, with the most unnatural having the worst score.

The problem faced, though, is that some specific examples were against the overall conclusions obtained from the tests that came before. The sentence with the lowest score obtained was "All small large cubes are large", with 46.1 of perplexity. Said score points it as less complex than the most simple sentence tested, which is wrong.

Therefore, we concluded that, although a good foundation for a metric, using Perplexity presented some problems in not being consistent and it is not good enough to measure Naturalness.

4.1.2. SLOR

Syntactic Log-Odds Ratio (SLOR) is defined as a naturalized language model score, for fluency evaluation of Natural Language Generation, presented by Kann et al. (2018). The formula can be seen below:

$$SLOR = \frac{\ln Probability\ (bigram)\ - \ln Probability(unigram)}{Length}$$

Although similar to Perplexity, it accounts for the unigram probability of every word. This means that, if a sentence presents a less usual word, it will not be heavily penalized, giving a bigger focus on the structure itself.

Considering the application in our project, its use was adapted. In the original publication, the probability was calculated using an LSTM (Long-Short Term Memory, an artificial recurring neural network), which requires a large corpus to train. Since our domain is more limited, it was decided to use the Language Model already trained.

The results obtained (Table 2) were not satisfactory, with no clear pattern emerging from the scores. Not even the patterns observed with Perplexity were observed here, indicating that the problem is probably in how SLOR was utilized.

Sentence	Perplexity Score	SLOR
“all small large cubes are large”	46.1	-1.83
“all cubes are red”	46.3	-2.22
“all cubes are red and all tetrahedra are red”	53.2	-1.98
“all red cubes and tetrahedra are large”	67.5	-2.32
“all large tetrahedra and cubes are cubes”	73.7	-2.17
“all cubes tetrahedra and cubes are cubes”	78.8	-1.89

Table 2 – Results of Perplexity and SLOR. The Perplexity Score overall increases with the more complex structures, with the outlier of the worst sentence being the lowest score, making it unsuitable for the application. The SLOR value does not present a clear pattern and, therefore, is not useful for this metric.

4.1.3. Readability Metrics

During the research for the Naturalness metric, it was discussed using a Readability Metric. It might be similar enough to the objective desired, so it was valid to test it out. Two were selected: Flesch Reading Ease and Automated Readability Index. The formulas can be seen below:

$$FRE = 206.835 - 1.015 * \frac{words}{sentences} - 0.846 * \frac{syllables}{words}$$

$$ARI = 4.71 * \frac{characters}{words} + 0.5 * \frac{words}{sentences} - 21.43$$

It is possible to observe that it is very reliant on the number of words, indicating that the length of the sentence influences the results. It should also be noted that the number of sentences is considered. These factors are not a problem when considering small texts, the case where these metrics are commonly used.

For a single sentence however, these metrics did not work. Its score was indicating which sentence was shorter, but not which was more readable.

Considering that the purpose of the tests was to check if it was a viable path to follow, the conclusion is that using such metrics for the current system is not useful unless they were heavily modified. For this, some separate work is required.

4.1.4. Conclusions

The methods tested did not achieve the results desired, with Perplexity coming closer. For future work, it is recommended to look for alternatives for Language Model or approach the problem differently.

One possible path to follow is to use Neural Language Models (such as BERT) as an evaluation of short sentences. Such work was already developed, as proven by the BERTScore (Zhang et al (2019)), an evaluator for text generation. This specific application is very computationally expensive, requiring GPU or cloud processing, which might be too powerful for single sentence evaluation. Therefore, further research might be required to determine alternatives.

4.2. Clarity

The initial idea was to use a Probabilistic Context-Free Grammar (PCFG), to have a probability for each sentence construction in the grammar. Since Tarski's World domain has a very specific set of words used, it was necessary to induce a PCFG from a set of sentences.

To do this, a set of sentences were obtained from the textbook "Language, Proof and Logic", by Dave Barker-Plummer, Jon Barwise, and John Etchemendy. It was a good source of sentences in this domain because it is the companion textbook for the Stanford University course that generated Tarski's World. In total, there are 241 sentences in this domain.

The next was to use this created database of sentences in a Natural Language Processing parser, to obtain the Parts of Speech (POS) of each word. For this process, it is necessary to use a parser already trained. There are a few available online but the one chosen was the CoreNLP Parser, developed by Stanford NLP Group. Run entirely online, it has to run a Java script parallel to its implementation in Python. Although it adds an extra step, its process is very fast and takes a few seconds to run everything.

To finally induce the PCFG, it was used the function "induce_PCFG" from the NLTK library. The output is grammar with probabilities, as expected. There are 349 productions for this database. This means that there are 349 POS associated with words and sentence structures.

Before moving on to the system, it is necessary to define what a sentence structure is. The sentence structures are the multiple possible ways of parsing a sentence. Theoretically, an unambiguous sentence would present only one structure, whereas an ambiguous one would present two (or more). The number of possible structures is determined by the complexity of the grammar since it dictates what possible combinations of POS exist. An example of the concept can be seen in Figure 3.

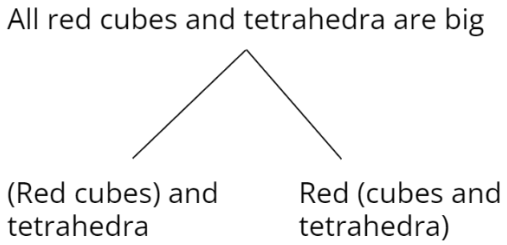

Figure 3 – An example of a sentence with multiple structures. In this case, the possible readings indicate a structure where "red" is only applied to "cubes" and a case where "red" is applied to both types of objects

The following step is using the proposed metric with the sentences generated. The initial test was using directly the grammar with a parser included in the NLTK library. It was observed then that the number of possible structures generated was too high for simple sentences. When analyzing the output, it was noted that nested POS were surrounding a single word, creating multiple structures. In practice, it did not change the syntactic structure of the sentence, but only over-specified the type of a single word.

For this reason, a new process was introduced after parsing the sentence. Since our main objective with this metric is to know the number of possible structures, the POS of each word is irrelevant. Therefore, the POS were substituted by "X" and the nested POS were substituted by a single "X". The results obtained then are solely focused on the possible general structures of the sentence. This cut the number of possible structures by a significant amount.

The results obtained from the first set of sentences were promising (Table 3). An unambiguous sentence like "all cubes are red" presents two possible structures, with the best one having a 99.98% probability over the other one; "all red cubes and tetrahedra are large" has 38 possible structures, and the best one has a 94.66% probability over the other ones. Although the difference in probability did not point to the desired direction, the number of structures did.

When testing with a larger number of sentences, though, it was observed that the metric did not behave in the way that was desired. A sentence like "all cubes are red and all tetrahedra are red" is very long, albeit not ambiguous. The parser pointed out 335 possible structures, which is supposed to indicate that a sentence is very ambiguous. When given the sentence "all small large cubes are large", a very incomprehensible sentence, it presented 12 structures, although the best one has a low probability of 51.45%. From the observations, then, it was possible to determine that the probability points out when a sentence construction is odd and that the number of structures is being affected by something else besides the ambiguity.

Sentence tested	Number of structures	Probability of the best structure
"all cubes are red"	2	99.8%
"all red cubes and tetrahedra are large"	38	94.66%
"all cubes are red and all tetrahedra are red"	335	95.3%
"all small large cubes are large"	12	51.45%
"all large tetrahedra and cubes are cubes"	38	94.66%
"all small cubes are red"	5	99.9%
"all large tetrahedra are cubes"	5	99.9%

Table 3 – Results obtained for the Clarity metric. When comparing sentences with the same number of structures, it is possible to observe that they have the same number of words, indicating that the result is associated with the length and not what was initially envisioned

Upon further testing with regular sentences with varying lengths, it was observed that the main factor for the number of possible structures is the length of the sentence. This means that the higher the number of words, the higher it is the number of structures. It was expected to be a factor, but its influence is too great. A possible reason for that is that the whole process of inducing a PCFG generates a grammar too complicated, to the point that the simplifications made affected the final result.

One way to solve it then is to write a grammar for this domain. Considering the fragment, it is an exhaustive but possible task. The only limitation of this approach is that the metric is as good as the grammar written. Therefore, all possible structures have to be delimited so they can work properly.

For future work, it is recommended to find an alternative between the induced PCFG and the manual grammar.

4.2.1. Conclusions

Although the metric showed promising initial results, further tests with complex cases made it evident that it was not reliable enough in its current form. To determine a possible problem, it is necessary to understand the way the metric works: an initial database is parsed through CoreNLP (an online pre-trained parser) and induces a grammar from the results. Upon further inspection, it was noted that there are more rules than necessary, with minor differences (or overlapping rules) creating multiple structures and, therefore, not providing an accurate result of actual different structures. For this reason, it might be necessary to write hand-made grammar.

To write the grammar, though, it is required to determine what possible sentences have to be included in the grammar. Since it includes the sentence and its possible structures, the number of variations could be infinite. Given that the objective is to use it in a metric to determine the best sentence generated, the grammar should include structures that are accepted as valid and clear. It is necessary then to do a survey and determine what are the interpretations accepted by readers in general and which are not.

5. Experiment: Perceived ambiguity of scopally ambiguous sentences

While working on the grammar for the Clarity metric, it became clear that it is necessary to determine what types of sentences are acceptable to be generated, especially regarding ambiguity. For this domain, it was observed that a common form of ambiguity is related to the scope of a word. For example, in a sentence like “red circles and squares” the word “red” can have two different scopes: it is a characteristic of just the circles (narrow scope); or for both squares and circles (wide-scope). This can affect the clarity of the sentence since there are multiple valid possible readings.

The main interest is to determine if there is a presence of a nocuous ambiguity. This means it affects the comprehension of the sentence to the point of becoming unclear what the desired meaning is. For this reason, sentences with this characteristic should be avoided in language generation and therefore it is interesting to determine if any of the selected sentence structures present it.

It should be pointed out that similar work was done, including the one presented in Section 2.2.2. Studying ambiguity is something pertinent for the field and necessary to determine a good sentence generation, as well as determining a clear sentence. What the current work strives to achieve is to examine a few examples common to the domain studied, with less focus on the implications at large and more on the application at hand.

Therefore, we decided to conduct a small survey to determine if said ambiguity is perceived as problematic to the comprehension of a sentence.

5.1. Questions

The main aspect being studied are the possible readings accepted by the participants and if there is an ambiguity generated by a specific element in the sentence.

The questions formulated for this study are as follows:

Question 1: Is there a preference between narrow and wide scope for predicates that indicate color? So if a sentence presents the fragment "red squares and circles", for example, it is necessary to determine if the participants accept both scopes or prefer one of them.

Question 2: Can the quantifier "all" be applied in narrow scope? For example, in a sentence like "all squares and circles", is it possible to interpret it as all squares and some circles?

Question 3: Does a sentence like "all squares and circles are red" (where there is no attribute to the objects) has a preferred reading?

Question 4: Does the presence of numbers in the sentence alter the perception of possible interpretations? So a sentence with the same structure but instead of a quantifier presents a number is perceived in different ways and do different numbers affect it too?

These are the main questions that were in mind when deciding the sentences used. Other observations were perceived but will be further discussed in the Results section. For now, these are the ones to keep in mind.

5.2. Design and Materials

Considering the questions, there are three sentence structures to analyze: sentence with attribute and quantifier; sentence with just quantifier; and sentence with numbers. Since the scope ambiguity happens with the presence of connectives, both conjunction and disjunction will be considered. It was also decided to include three different numbers, to observe if it affects the perception of the participants.

Given that the ambiguities analyzed are presented only in the first part of the sentence fragment (before the "are"), we decided to maintain constant the second part, with it being "are on the left". Said decision was made to create an image that can easily be discerned, with the important part to parse being the first part of the sentence.

We decide to test also if the position of the color affects the interpretation and how the scope is perceived. The way to write the sentence "all red squares and circles" then becomes "all squares and circles that are red". It is important to note that the narrow scope indicates now the other object, but the image presented is the same. This way, we should observe if there will be any significant alteration.

A selection of 12 sentences was formulated, with 3 to 4 interpretations presented for each one, including a control one. The selection is presented below.

- "All red squares and circles are on the left" (and variation phrased as "… circles that are red…")
- "All red squares or circles are on the left" (and variation phrased as "… circles that are red…")
- "All squares and circles are on the left"

- “All squares or circles are on the left”
- “X squares and circles are on the left” (X being either 3, 5 or 8)
- “X squares or circles are on the left” (x being either 3, 5 or 8)

To evaluate each interpretation, we agreed that the best way to present them is through visual means. An example can be seen in Figure 4. This initial design clearly defines what is left and right, with a line in the middle. The downside presented, though, is that it can take a while to figure out what the interpretation means. For this reason, another possible representation discussed was using a text description of the interpretation. An example can be seen in Figure 5.

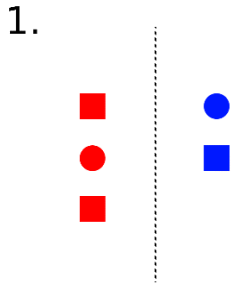

Figure 4 – Initial representation of an interpretation. It has line in the middle to clearly indicate what is left and right.

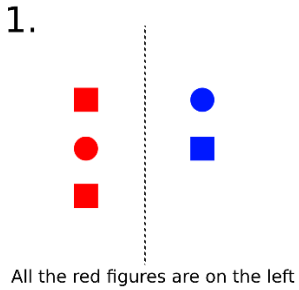

Figure 5 – Representation with description.

To determine whether the text description was necessary or not, a small group of people was presented with a selection of 4 questions, with one alternating pair presenting images with descriptions and the other presenting just images. Afterward, they were asked which they preferred and 66% said that an image with a description was clearer. When observing the results, it was also observed that, with this presentation, the participants were more certain about their opinion.

For each question, all the interpretations are initially presented, as in Figure 6. Then in each evaluation, the image of the specific interpretation is shown again, like in Figure 5. All images are available in https://github.com/igorkmi/Experiment-Perceived-ambiguity-of-scopally-ambiguous-sentences .

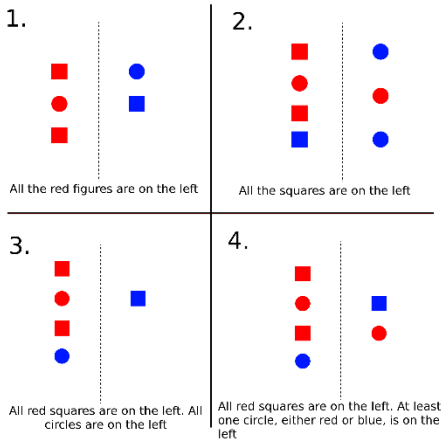

Figure 6 – Presentation of all the interpretations in a single image.

5.3. Procedure and Participants

The participant, before answering the questions, were tasked with answering a few personal questions: level of English (according to the proficiency scale); experience with logic; native language; and if they present monochromacy (complete color blindness).

The survey was made with the presentation of the sentence, the image of all the interpretations, and asking how well each interpretation fits the original sentence. The evaluation was made using a 4-point scale, going from “Doesn’t fit” to “Perfect fit”, with no central value so the participant has to decide between negative or positive feedback. Each evaluation is independent, thus a participant can rate all interpretations with the same value, allowing them to indicate that both interpretations seem equally likely or unlikely.

As a means to minimize biases, the order of the questions was scrambled, preventing the clustering of similar sentences. This way, the participants would have to deal with different types of sentences before having another example of the same type. The order of the questions can be seen below:

- All red squares and circles
- 8 squares or circles
- All squares or circles
- 5 squares and circles
- All squares or circles that are red are on the left
- 3 squares and circles
- All squares and circles
- 5 squares or circles
- All squares and circles that are red are on the left
- 3 squares or circles
- All red squares or circles
- 8 squares and circles

The order of the interpretations for the number of sentences was also shuffled. This way, the order in which the participants have to evaluate them is different, preventing receiving the same answers if the participant remembers the question of the same type.

The survey was done entirely online, using Google Forms. This allowed me to contact people from multiple backgrounds and countries and made it easier to make adaptations and create analytics.

Since this is a pilot study, there were 20 participants, with none presenting monochromacy. The native language of the participants is very varied, with the presence of Portuguese, French, Dutch, and Greek as the main ones. Regarding the level of English, 75% identified as above B2 (Upper Intermediate Level) and none identified as A2 (Basic Level). 60% claim to have either no experience or little experience in logic, with the minority of 15% claiming to be very experienced. The information can be seen in Figures 7 and 8.

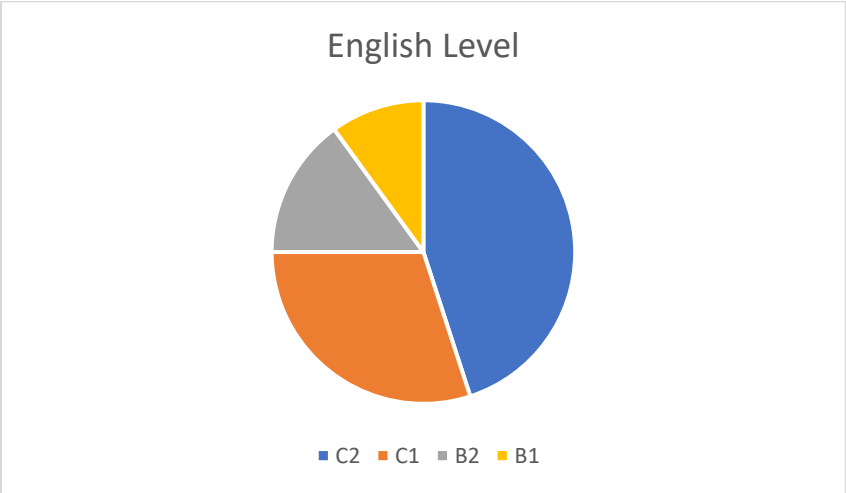

Figure 7 – Distribution of English level from participants

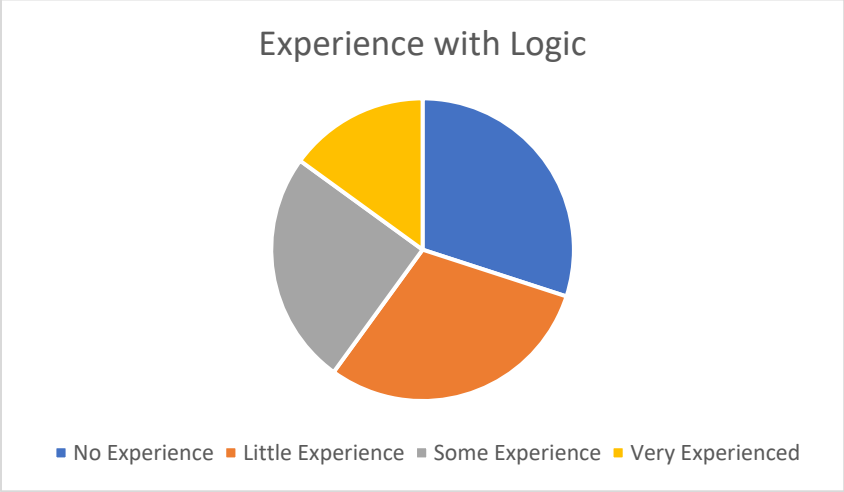

Figure 8 – Distribution of Logic Experience from participants

From this information, it is possible to determine that the participants have a good to great level of English and that, given the varied background, not everyone has a background that involves logic. Considering this, the results can be extrapolated to an ample group of English speakers.

5.4. Results

The first component analyzed was the impact of the level of English and experience in logic in the results. As mentioned in the previous section, the majority of the participants have an above-average comprehension of the language. This did not impact in any significant way the performance of the participants, with the ones below average having similar results to the ones with the best level of English.

Regarding the experience in logic, there were no conclusive results of it correlating with the way the participants perceive the interpretations. When comparing the participants with no experience to the ones very experienced, the results were not different enough to conclude anything. It was even noted that, in some cases, some very experienced failed the control interpretation whereas some not experienced evaluated it correctly.

To observe the answers to the questions, it is necessary to analyze each sentence (or group of sentences) separately, going through the list presented earlier. It is also important to notice that not all images will be presented here since the number of them is too extensive. For this reason, they will be available in https://github.com/igorkmi/Experiment-Perceived-ambiguity-of-scopally-ambiguous-sentences , for consultation.

5.4.1. "All red squares and circles are on the left"

The sentence "all red squares and circles are on the left" has the elements necessary to study whether questions 1 and 2 are valid.

Regarding the scope of "all", when presented with the interpretation that some circles are on the left (meaning that "all" has narrow scope) 75% of the participants rated it negatively (with 80% of them rating it as "Doesn't fit"). This points to the narrow scope of "all" not being accepted as valid (question 2).

For the scope of "red", the response is as expected: both scopes are the best-evaluated interpretations. There is a small difference in percentage, with the narrow scope having 90% evaluating it positively and the wide scope having 85%. It was interesting to notice, though, that the wide scope has more participants evaluating it as the highest value, while the narrow scope has a bigger range of different evaluations.

To determine if there is no clear preference, a comparison between both interpretations was made, as presented in Figure 9. It becomes clear that the majority of the participants rated both with the same value, indicating that there is no preference. All of this considered, it indicates that there is no clear preference for the narrow or the wide scope of color predicates (question 1).

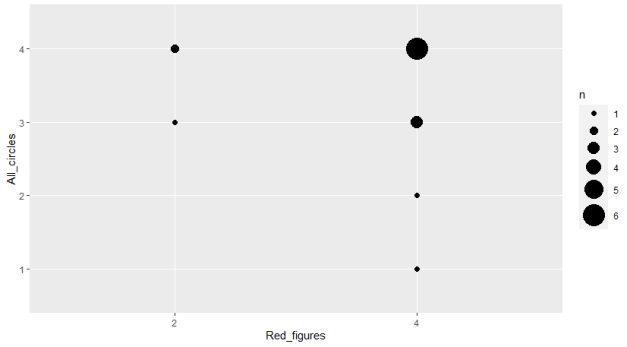

Figure 9 – Comparison between the interpretations with "red" having narrow scope (All_circles) and wide scope (Red_figures). It is clear that the majority of the participants rated with the best value: 4.

5.4.2. "All squares and circles that are red are on the left"

Comparing the results of "all squares and circles that are red are on the left" with the previous sentence, it was noticed that it had a very similar result, with the same interpretations being accepted and rejected. The biggest difference was in the proportions of each evaluation. The interpretation that indicates narrow scope for "all" had 90% of the candidates rating it negatively. For the interpretation that considers red as wide scope, the positive evaluations raised to 95% and for the narrow scope, it lowered to 80%, inverting the order of the previous sentence. The conclusions regarding questions 1 and 2 are the same.

5.4.3. "All red squares or circles are on the left"

The sentence "all red squares or circles are on the left" has the elements necessary to study whether questions 1 and 2 are valid.

When presented with the interpretation with narrow scope for "all", the participants rejected it, with only 10% evaluating it as positive (but none evaluating it as a "Perfect fit"). This points to the narrow scope of "all" not being accepted as valid (question 2).

Regarding the interpretations that deal with the scope of "red", the response was mixed, since they were both accepted by almost half the participants: narrow scope was positively rated by 55% of them, and the wide scope by 60%. It should also be noted that the narrow scope interpretation was the most divisive of the survey, with each value of the evaluation ranging from 20 to 30% of the participants.

When comparing the evaluations of both interpretations, it became clear that the participants did not agree on which one was better: the people that evaluated "red" with wide scope as a 4 mostly evaluated the narrow scope as a 3 or 4 (with a few cases of 1 and 2 too). Interestingly, the biggest concentration of the comparison is with both interpretations rated as 1. The comparison can be seen in Figure 10. All of this considered, it indicates that there is no clear preference for the narrow or the wide scope of color predicates (question 1).

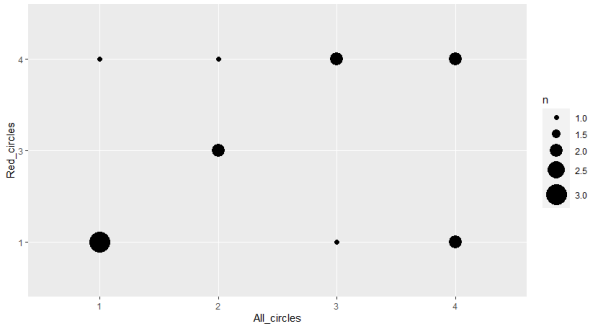

Figure 10 - Comparison between the interpretations with "red" having narrow scope (All_circles) and wide scope (Red_figures).

5.4.4. "All squares or circles that are red are on the left"

Comparing the results to the previous sentence, they are similar, but with alterations to the proportions. The interpretation of the narrow scope for "all" was rejected again, and the top interpretations remain the same. This time, though, the gap between them was bigger. The interpretation of "red" as the wide scope was accepted by 70% of the participants, while the other was accepted by 55%. It was also noted that a considerate number of candidates (40%) failed to control. This indicates that this sentence structure was problematic to understand. The conclusions regarding questions 1 and 2 are the same.

5.4.5. "All squares and circles are on the left"

The sentence "all squares and circles are on the left" has the elements necessary to study whether questions 2 and 3 are valid.

When presented with the interpretation with "all" as narrow scope, it got rejected by 70% of the participants. Although better than the results observed in the previous questions, it still points to the narrow scope of "all" not being accepted as valid (question 2).

There was one preferred reading out of all presented, with all the participants rating it positively. This indicates a positive answer to question 3.

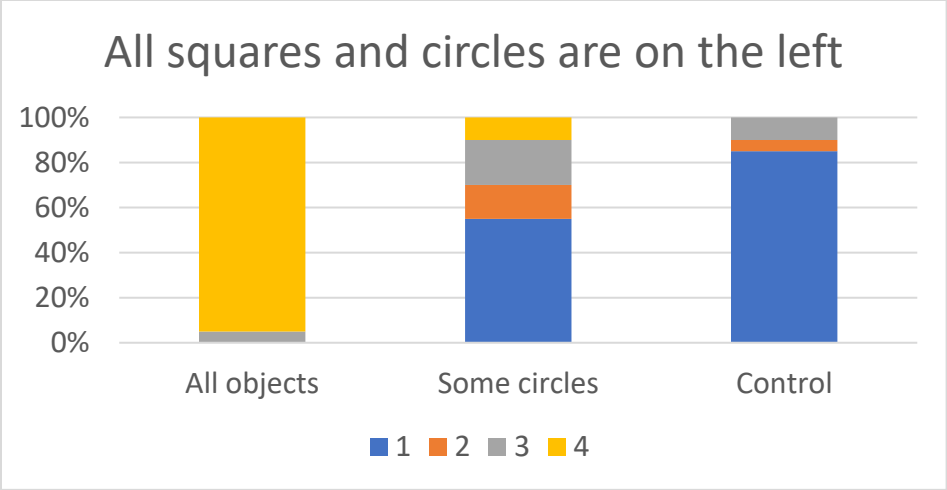

Figure 11 - Results of the sentence "All squares and circles are on the left". The preferred reading was the wide scope, with narrow scope having a small number of participants positive to it.

5.4.6. "All squares or circles are on the left"

The sentence "all squares or circles are on the left" has the elements necessary to study whether questions 2 and 3 are valid.

Similar to the previous sentence, the interpretation that posits that "all" has a narrow scope was rejected, with 95% rating it negatively. This indicates a negative answer to question 2.

Also, like the previous sentence, there was only one reading accepted by the majority of the participants, with only 15% rating it negatively. This points to a positive answer to question 3.

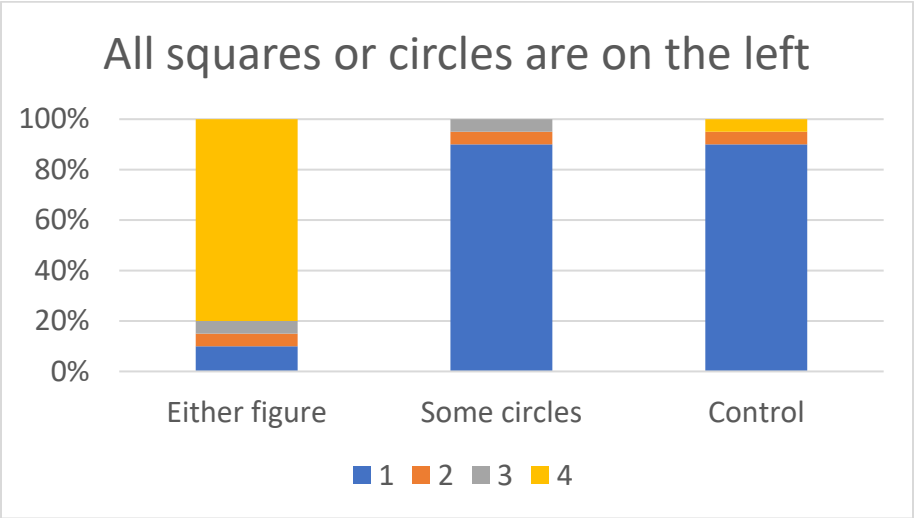

Figure 12 - Results of the sentence "All squares or circles are on the left". The preferred reading was the wide scope, with narrow scope having a small number of participants positive to it. When compared to the previous sentence, the amount of participants positive to the second-best reading was even smaller.

5.4.7. "X squares and circles are on the left"

For the sentences with numbers, they will be analyzed comparatively with each other, since the impact of the number is the most interesting aspect. It is also important to point out that they all present the same interpretations, with the only variation being the number of objects in the image.

The possible interpretations for this type of sentence are the number having a wide scope; the number having a narrow scope; the number representing the sum of both objects presented. The results of all the sentences can be seen in Figures 11 through 13.

Overall, the most accepted interpretation was with the number having a wide scope, with the highest rating being 90% (when the number is 5) and the lowest being 75% (when the number is 8). Considering that it is the easiest reading to make, it was expected. Therefore, the main focus is on the variation between the percentages for each number: it seems like the confidence of this being a valid interpretation diminished throughout the survey since the highest one was the first to appear and the lowest was the last. It still gives a positive result but indicates that some participants had doubts, especially with a higher number of objects presented.

Interestingly, there was no prevalence of a second-best interpretation: for the number 5, sum prevailed; for the number 3, they were tied; and for the number 8, the narrow scope was better. What is very curious is how the ratings vary between the numbers.

The interpretation of narrow scope, although it has a good amount of positive ratings, the majority of them are partially positive (3 out of 4). This means that the number of participants that thought it completely fit was few, varying from 27% to 35% of the total positive ratings.

On the other hand, the interpretation of the sum varied in the total amount of positive ratings but got consistent "Perfect fit" ratings, ranging from 58% to 77% of positive ratings. So, even though it was not consistently better than the narrow scope interpretation, more participants thought that it was an interpretation that fit perfectly.

So the reason the narrow scope eventually got more positive ratings was not that the participants diminished their perception of the interpretation of the sum, but because more of them rated the narrow scope as partially positive. One possible explanation for that is that returning to the same problem made it sound more plausible.

When considering the numbers in sequence, there were small details that coincided: the sum interpretation got slightly better ratings with the higher numbers; the participants struggled more with the control interpretation in the higher numbers (albeit few failed). With all the present information, there is some indication that the presence of numbers in the sentence (and how high it was) affected the perception of some interpretations, albeit not strong enough to give a definitive conclusion about it (question 4).

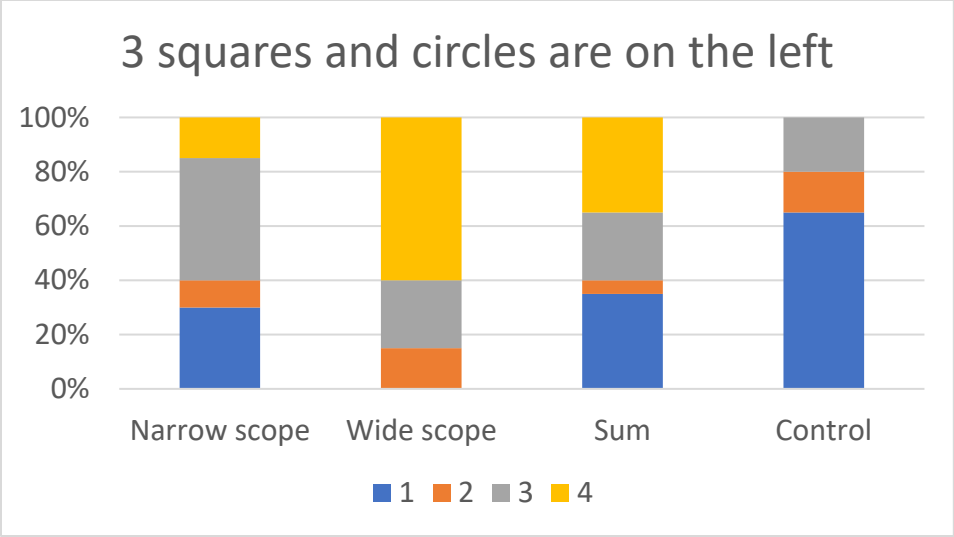

Figure 13 – Results of the sentence "3 squares and circles are on the left". The preferred reading was the wide scope, with narrow scope and sum very close in second. It has the least amount of people failing control in the category.

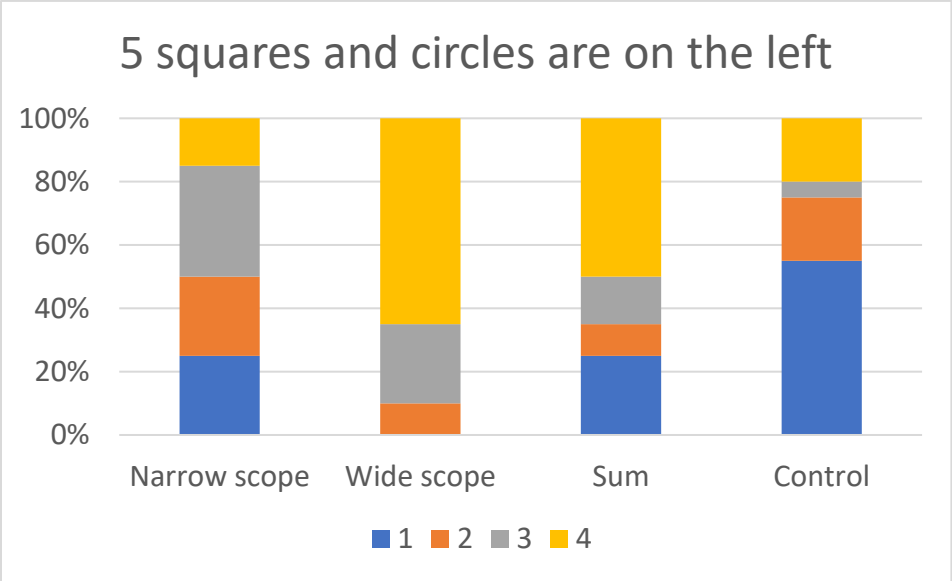

Figure 14 – Results of the sentence "5 squares and circles are on the left". The preferred reading was the wide scope, followed by sum.

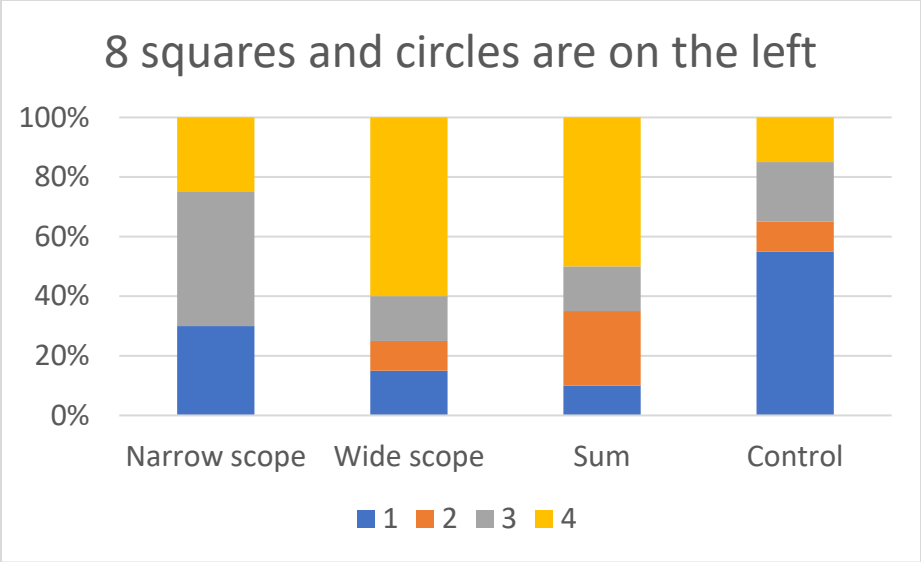

Figure 15 – Results of the sentence "8 squares and circles are on the left". The preferred reading was sum, followed by wide scope.

5.4.8. "X squares or circles are on the left"

For this set of questions, there were few possible interpretations: the number is attributed to an object, and the number represents the sum of the objects. The latter might seem like an unlikely interpretation, but it can be used in informal settings when the interlocutor means X number of objects that can be either squares or circles, independent of the proportion. It is a very similar idea to the sum interpretation of the previous sentence, but with a phrasing that appears generally only in spoken English. The results can be seen in Figures 14 through 16.

As expected, the main interpretation is the one with the number attributed to only one of the objects. There are just a few outliers, but nothing noteworthy since the minimum positive rating was 95%.

For the sum interpretation, the results were surprising: it was generally accepted, with the positive rating ranging from 65% to 75%. This means that, although not a common interpretation of the sentence, it was still considered valid.

One thing to notice, though, is that there was no pattern emerging from the order that which the participants answered the questions. Therefore, the changes made from going from the first number (8) to the second (5) did not carry between the second (5) and the last number (3). When comparing the results of 8 and 3, the number of positive ratings for the sum interpretation did not change, but the share that is partially positive increased; the share that was partially negative decreased too, therefore pointing to the participants not being certain about the interpretation. The number 5 points to a somewhat different trend: the partial values increase, indicating convergence to the middle point.

For us, this indicates that the number could have influenced the perspective of the participants, a positive answer to question 4.

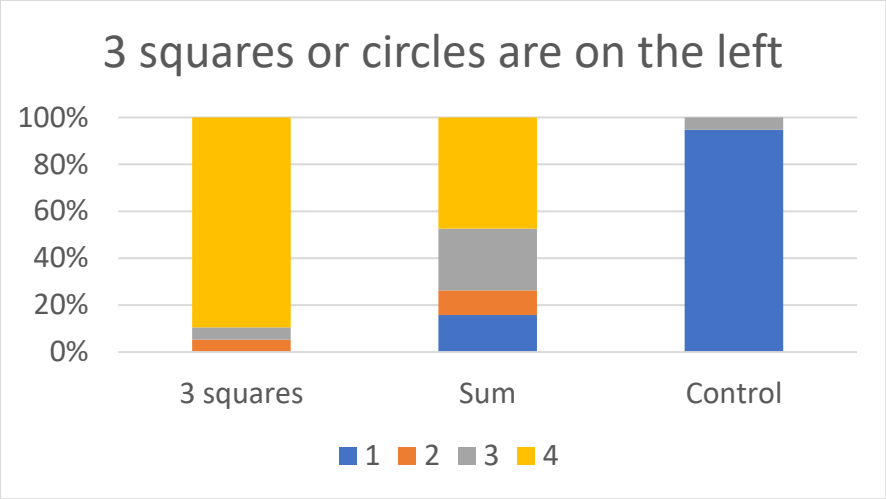

Figure 16 – Results of the sentence “3 squares or circles are on the left”. The margin of difference between interpretations is smaller than the other numbers.

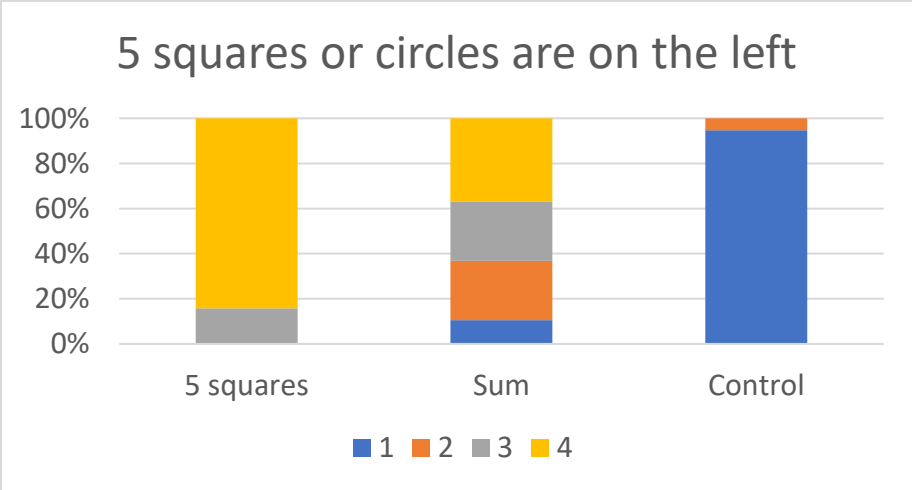

Figure 17 – Results of the sentence “5 squares or circles are on the left”. This number had the worst results for the sum interpretation.

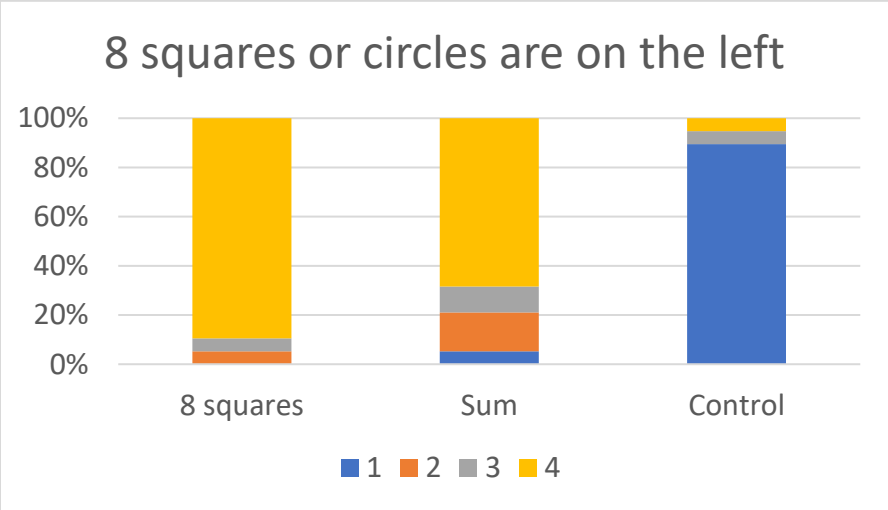

Figure 18 – Results of the sentence “8 squares or circles are on the left”. This number has the lowest amount of negative ratings for sum.

5.5. Conclusions from the experiment

From the results obtained in this pilot study, it was possible to conclude a few things, based on the initial questions: a color predicate can be interpreted as narrow and wide scope, without preference; the quantifier “all” cannot be applied in a narrow scope, and a sentence with no attributes has a clear meaning.

Regarding the question that different numbers affect the perception of a given sentence, the data collection points to a variation between each evaluation but it is not clear whether it is solely based on this factor or on the way the experiment was conducted. There is some indication that it can affect (since the results were not completely consistent) but it is not strong enough to conclude. For this reason, we suggest that further research should be made on the matter.

Concerning the utility of the system, it is necessary to observe through a higher-level perspective. From the study, it was possible to conclude that: a sentence with no adjective, regardless of whether it is a conjunction or a disjunction, is unambiguous; a sentence with an adjective, regardless of whether it is a conjunction or a disjunction, is ambiguous enough to confuse; changing the position of the adjective alone did not affect significantly the ambiguity perceived; a sentence with a number present in its structure is perceived as ambiguous. It should be noted that, even with the presence of ambiguities, some cases have a main interpretation. It should be decided, then, if it is better to have a longer sentence in favor of clarity or if it is acceptable to have this ambiguity given the propensity of being the accepted reading.

When comparing with the results obtained using the Clarity metric, there were some similarities: the use of conjunction or disjunction did not affect the value of it, and a sentence with no adjective got a better result than one with an adjective. The results differ when considering the other possible structures: changing the position of the adjective changed significantly the value of the metric (while in the survey, there was no significant difference); and a sentence with a number present was better evaluated than all the other ones, while it was considered ambiguous in the survey. From this, it is possible to observe a dissonance between the expected results and what was obtained through the metric, confirming that it needs to be reworked.

From the results of this experiment, scopally ambiguous sentences involving adjectives can be indicated as having nocuous ambiguity. All the sentences that presented this format had the same problem in clarity and difficulty in determining a prevalent reading, defining them as not clear enough to use in a language generation system. When a scopally ambiguous sentence was presented but using instead a number, the ambiguity was enough to split the perception of the possible reading, but it was still possible to observe a preferred one by a small margin.

It is also important to indicate that, since it was a pilot study, the number of participants is small and the conclusions made should be considered as strong indications but not irrefutable proof that it is true. For that, research with a considerable population would be required, which is beyond the scope of this project. Since our interest was in the way that the sentence is constructed, the results obtained were satisfactory.

6. Conclusions

As established in the subsection Previous Work, there were some efforts into translating from FOL to Natural Language, with varying levels of complexity and methods.

Hatzilygeroudis and Mpagouli (2007) created an interesting sentence fragment and was a good first step in the effort; Flickinger, D. (2016) created a very robust system that allows for multiple sentence

constructions, with the limitation of not including quantifiers. With this background, the present work contributes a bit by building upon the structure of the former and removing the limitation of the latter.

The algorithm developed is a good first step into this type of Language Generation: the sentences are comprehensible and it fulfills the translation of the desired sentence fragment. For future expansions, it is necessary to develop a way to deal with global negations and how to expand to other domains.

Considering the metrics, the overall results were mixed: there were no definitive metrics found for either Clarity or Naturalness, but the exploration of them led to possible directions to continue working on.

For Clarity, the metric had some flaws that could be the product of using induced grammar. To solve that, it is either necessary to create a manual grammar or find a better method of inducing it.

For Naturalness, the metric did not work as intended, given that the length of the sentence affected all the different formulas that were tested. Therefore, it is recommended to determine if there is a way to use Language Models with a very diminished effect on the sentence length; or follow through the path of Neural Language Models, although more computationally expensive.

Regarding the survey, it was possible to determine a few sentence structures that can be used in sentence generation. In Kahn, van Deemter (2012), the authors pointed out that the strength of the adjective in an ambiguous sentence can determine whether it has a narrow or wide scope. Through the survey, it was indicated that the color adjective has no clear scope and therefore is not strong enough (or weak enough) to have a determinate scope. Considering the scope of the quantifier “all”, it was indicated that the simplest reading was the preferred one, even if it is possible to interpret it as narrow scope for some. This shows that this sentence is, for sentence generation, not ambiguous. Finally, in the sentences with numbers present, the results did not lead to a conclusion on whether they affect comprehension or not and therefore should be further investigated in future works.

In conclusion, the current effort to translate from FOL to English showed promising results, while investigating multiple methodological directions and possibilities. Some topics were more exploratory but overall indicate interesting directions for future work.

7. References

Barwise, J., & Etchemendy, J. (1993). Tarski's world. Stanford, Calif: CSLI Publ.

Flickinger, D. (2016). *Generating English paraphrases from logic*. From Semantics to Dialectometry.

Hatzilygeroudis I., Mpagouli A. (2007). *Converting first order logic into natural language: A first level approach*.

Kann, K., Rothe, S., & Filippova, K. (2018). *Sentence-Level Fluency Evaluation: References Help, But Can Be Spared!*. arXiv preprint arXiv:1809.08731.

Karttunen, Lauri. (1989) *Translating from English to Logic in Tarski's World*. Journal of Information Science and Engineering 5, pages 323-348

Khan, I. H., van Deemter, K., & Ritchie, G. (2012). *Managing ambiguity in reference generation: the role of surface structure*. Topics in cognitive science, 4(2), 211–231.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Naval Technical Training Command Millington TN Research Branch.

Mayn, A., & van Deemter, K. (2020). *Towards Generating Effective Explanations of Logical Formulas: Challenges and Strategies*. In 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (pp. 39-43).

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019*). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.*