Master Thesis

Modeling Alternative Sets with Recurrent Neural Networks

Guido Marvel Linders

Supervisors

Dr. Roberto Zamparelli
Prof. Dr. Matthew W. Crocker

Submitted in partial fulfillment of the requirements for:
Master's degree in Cognitive Science
Master of Science in Language Science and Technology

University of Trento
Saarland University

UNIVERSITÀ DEGLI STUDI DI TRENTO
CIMeC - Center for Mind/Brain Sciences

UNIVERSITÄT
DES
SAARLANDES

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged.

Place: Tilburg, The Netherlands

Date: April 3, 2020

Signature:

Abstract

Alternatives are things we could have said or heard, but did not hear. Their implicit nature makes them hard to study from a cognitive point of view. They, however, play a central role in the compositional meaning of sentences containing focus, negation and other phenomena. In this thesis we explore the possibility of modeling alternatives with language models. Because alternative sets vary widely in their size and contents, we use focus particles to constrain these sets. Focus particles place an element in focus and also restrict the set of alternatives of the elements they operate on. We investigate the extent to which a simple long short-term memory (LSTM) and a simple bidirectional long short-term memory (BiLSTM) can capture human intuitions on alternative sets through the focus particles *even* and *only* and using two evaluation metrics: syntactic log-odds ratio (SLOR) and perplexity. Our experimental setup is divided in two parts. In the first part we collect a corpus of simple sentences containing either *even* or *only*. We remove the focus particle from each sentence and collect human judgments on which focus particle is the most plausible in the given sentence. Next, we evaluate the models on these sentences and compare the human judgments to the model's judgments on three factors: the naturalness of the sentences, the contribution of the focus particle to a sentence and the predictions of the most plausible focus particle in a sentence in a decision task. We show that the LSTM model is able to capture some human intuitions on focus particles, especially when the focus particle is present in object position. We furthermore show that perplexity is not a good evaluation metric for this task, since the models are sensitive to frequency biases. In the second part we artificially construct sentences of the form *Subject Verb Object and even* ___. and collect alternatives from human participants. We compare the human generated alternatives on their likeliness or unexpectedness. Here, however, the results show no consistent evidence that the models are able to capture the human intuitions on the ordering of the alternatives on a scale. Overall we conclude that the ability of the language models to learn about alternative sets is limited, but not completely absent. Further research is needed to get a better understanding of the extent to which language models are able to learn about alternative sets.

Acknowledgements

First of all, I would like to thank my supervisor Roberto Zamparelli, who has guided me from the point where I had no idea what topic to do my thesis on to the completion of this thesis. His comments, steering and support have always been very valuable. I would also like to thank Matthew Crocker who has helped me on the side of Saarland University.

Secondly, I want to express my gratitude to Shammur Chowdhury for providing me with the computational language models that were used in this thesis. She helped me with understanding the models and taught me how to use them.

Also, I want to thank the LCT consortium for providing me with the opportunity of living and studying in two different countries. In particular, I want to thank Bobbye Pernice who was always available to help with organizational problems.

I furthermore would like to thank my friends I made through university and also outside university in both Germany and Italy and that made this journey one to never forget.

Finally, I have to thank my parents who supported me in every possible way throughout this journey.

Contents

Chapter 1

Introduction

Alternatives are things that a speaker could have said, but did not, or a listener could have considered, but did not hear. Alternatives play a very important role in the compositional meaning of sentences containing focus, negation and other phenomena, but their implicit nature makes them hard to study in a formal manner. Alternatives are very apparent in language. Consider (1), where one could consider the alternative that John did not go to the party, but instead went to see a movie.

(1) John went to the party.

However, the alternative that John went to the toilet is not a very likely alternative in this sentence, as can be shown by considering some possible continuations as in (2).

(2) a. John went to the party and not the movie.

b. ??John went to the party and not to the toilet.

c. ??John went to the party and not on a Himalaya expedition.

It is not clear how the set of plausible alternatives is restricted and which factors make some alternatives more plausible than others.

This thesis builds upon three different lines of motivation. The first line of motivation comes from the appeal to generate alternatives formally or computationally and the interest in making alternatives more explicit. Knowing which alternatives are considered can give us insight in how humans process language and ultimately tell us more about how the human mind works. Generating alternatives makes them explicit and makes us able to study them more easily. Understanding alternatives could furthermore help us in understanding many linguistic phenomena, such as focus, negation and also generic sentences. Experimental research on alternatives is present, but is limited in its progress by the implicit nature of alternatives (Chemla and Singh, 2014). In addition, formal models often lack a theory of how to restrict the set of alternatives and order them on plausibility (Fălăuş, 2013). Computational models to generate alternatives could therefore be a good alternative. To date, there have only been a few studies that

tried to computationally generate alternatives (Aina, 2017; Kruszewski et al., 2016; Storozum, 2018).

The second line of motivation stems from the linguistic topic of focus. Focus is a way to put emphasis on a particular part of an utterance and by this, induce alternatives. Focus particles are used to restrict this set of alternatives. By comparing sets of alternatives that were restricted using different focus particles, we could get insights into alternative sets in a more systematic way. Focus particles, thus far, have also received very little computational treatment, most likely because focus is not marked in written texts and thus not easy to represent in a model. The exact behavior of focus particles and the way in which they restrict a set of alternatives is not clear either and there is a lot of debate on their contribution, both in terms of general properties as well as on individual particles. That focus and focus particles play an important role in language is clear, however (König, 1991).

The third line of motivation behind this thesis stems from research on traditional language models (LMs). Recently these models have shown to be able to capture grammaticality and even acceptability judgments. In addition, distributional Semantic (DS) models turned out to be good at capturing alternatives (Aina, 2017; Kruszewski et al., 2016; Storozum, 2018). The question we address here is whether the traditionally more syntactic-oriented LMs are also able to capture alternatives, since they learn patterns in a more or less similar fashion as distributional approaches by representing words based on the surrounding context.

This thesis is exploratory in its nature and consists of two parts. In the first part we investigate the ability of LMs to capture human intuitions on the focus particles *even* and *only*. As we will see in Chapter 2, the semantic contribution of *even* and *only* is limited to their operation on the set of alternatives. Besides, they are very frequently used in language and show roughly opposite behavior (Greenberg, 2018, 2019). If these models are able to capture these intuitions, it would be evidence that LMs are able to learn some information about alternatives, since the contribution of focus particles is to modify the set alternatives. For this we will need to collect human judgments on focus particles and compare these judgments to judgments made by the LMs.

In the second part of this thesis, we investigate the ability of LMs to generate alternatives with the same intuitions as humans, using the scalar property of *even*. *Even* ranks the alternatives on some kind of scale. We create a setup where human participants provide us with alternatives that are ideally more surprising or informative than a comparison alternative. We investigate whether LMs can capture these scalar effects and what the influence of *even* is on the order of the alternatives that LMs generate. If LMs can capture human intuitions on these scalar effects induced by *even*, it would be further evidence that LMs are able to learn some humans intuitions on alternative sets.

This thesis is organized as follows: Chapter 2 gives an overview of the literature on alternatives and focus particles from a theoretical, experimental and computational perspective. We build a theoretical framework that forms the basis of our experimental setup. Chapter 3 outlines the experimental setup. We explain how we collected data

and it was used in human judgment tasks. We also explain which models and evaluation metrics were used and finally how we extracted alternatives from the models. Chapter 4 presents and discusses the results. Finally, Chapter 5 concludes this thesis.

Chapter 2

Theoretical Framework

This chapter outlines the relevant theory on alternatives and focus particles and also computational models that can be used to model the behavior of them. The foundations in this chapter are based on theoretical, experimental as well as computational research. We build a theoretical framework that will form the basis of our experimental setup. This chapter starts by introducing alternatives from different angles. We then move on to build the foundations of focus particles, before we move on to explain language models (LMs) and why they might be suitable for modeling alternatives.

2.1 Alternative Sets

The way alternatives are induced and constructed has been studied from different angles. There is large body of theoretical research that has tried to build formal frameworks that explain how alternatives are constructed. Experimental research has mainly studied the restrictions that are imposed on these sets. More recently, there has been computational research that showed it is possible to model alternatives, using Distributional Semantics (DS) (Aina, 2017; Kruszewski et al., 2016; Storozum, 2018). DS offers a way to represent the meaning of words through their distribution in context.

Each angle will be briefly discussed, starting with outlining the theoretical foundations, then moving to experimental findings before discussing computational approaches to alternatives.

2.1.1 Formal Basis

On the theoretical side, there are two main frameworks that try to explain how alternatives are formed from a formal point of view: Hamblin Semantics and Alternative Semantics.[1] Both frameworks will be briefly discussed. For a more comprehensive

[1] More recently a third framework has been proposed, which will not be discussed here. Büring (2016) introduced a framework that is able to derive alternatives, using mainly syntactic information and only little semantic and contextual information. First, using a few simple syntactic rules the impossible alternatives that cannot be combined using function application are excluded. In a second step, contextual restrictions are applied in a similar fashion as in the Alternative Semantics framework.

overview of the two frameworks, we refer the reader to (Fălăuş, 2013). Finally, we will cover theories on how and when alternatives are computed.

Hamblin Semantics

Hamblin (1973) proposed to treat a question as a set of propositions with possible answers to that question. Hamblin achieves this within the Montague grammar by replacing denotations by denotation sets, also called alternative sets. Every lexical entry is considered a set, with standard lexical entries being singleton sets and question words being sets with multiple entries. Building up phrases and sentences is done through pointwise function application. This is illustrated in (3) and (4). The examples are loosely based on examples in (Fălăuş, 2013).

(3) a. Ben laughed.

b. $[[\text{Ben}]] = \{\text{ben}\}$

c. $[[\text{laughed}]] = \{x : x \text{ laughed}\}$

d. $[[\text{ben laughed}]] = \{\text{laughed(ben)}\}$

(4) a. Who laughed?

b. $[[\text{who}]] = \{x : x \text{ is a person}\} = \{\text{adrian}, \text{ben}, \text{claire}, \ldots\}$

c. $[[\text{who laughed}]] = [[\text{laughed}]]([[\text{who}]]) = \text{laughed}(\{\text{adrian}, \text{ben}, \text{claire}, \ldots\})$
$= \{\text{laughed(adrian)}, \text{laughed(ben)}, \text{laughed(claire)}, \ldots\}$

In (3) we can see that the lexical entry *ben* is represented by a singleton set and the only element in the set is applied to the function, which is the verb *laughed*. In (4) we can see that *who* is represented as the set of persons. Each element in this set is applied to the verb in a pointwise manner. The result is a set of propositions that are potential answers to the question: *Who laughed?*.

A mechanism on how to decide and restrict the set of alternatives is not provided in the standard version of this framework. There has been debate on whether the set of possible answers to a question should be restricted to the set of true answers (Karttunen, 1977) or the set of all possible answers (Groenendijk and Stokhof, 1984; Hamblin, 1973).

Alternative Semantics

Alternative Semantics is a framework that was designed to explain focus in sentences (Rooth, 1985, 1992). Rooth (1985, 1992) proposed that focus evokes alternatives and in his framework each proposition has two semantic values. A standard semantic value and a so-called *focus semantic value*. The focus semantic value of elements that are not in focus is the singleton set containing the standard semantic denotation. The focus

semantic value of an element in focus is a set of contextually unrestricted alternatives of the same semantic type. This is illustrated in (5). Focus is induced by putting prosodic prominence on an element. This is called F-marking (Jackendoff, 1972). The element in focus is marked by an F. The example is loosely based on an example by (Fălăuş, 2013).

(5) a. $[\text{Ben}]_F$ likes ice cream

b. $[[\text{Ben}_\text{F}]]^o = \text{ben}$

c. $[[\text{ice cream}]]^o = \text{ice\_cream}$

d. $[[\text{likes}]]^o = \lambda y \lambda x.\text{likes}(x, y)$

e. $[[\text{likes}]]^o([[\text{ice cream}]]^o)([[\text{Ben}_\text{F}]]^o) = \text{likes}(\text{ben}, \text{ice\_cream})$

f. $[[\text{ben}_\text{F}]]^f = \{\text{adrian}, \text{ben}, \text{claire}, \ldots\} = A_e$

g. $[[\text{ice cream}]]^f = \{\text{ice\_cream}\}$

h. $[[\text{likes}]]^f = \{\lambda y \lambda x.\text{likes}(x, y)\}$

i. $[[\text{likes}]]^f([[\text{ice cream}]]^f)([[\text{Ben}_\text{F}]]^f) = \{\text{likes}(x, \text{ice\_cream}) \mid x \in A_e\}$

In the example, *Ben* is the element in focus. Interpretations with respect to the standard semantic value and focus semantic value are denoted by o and f, respectively. The set of alternatives is denoted by A and is of type e in this example, since *Ben* is an individual. In (5i), like in Hamblin Semantics, we apply pointwise function application and apply every element in the set of alternatives to the constituent *likes ice cream*.

Similar to the Hamblin Semantics framework, this framework also does not provide a mechanism on how to determine and constrain the set of alternatives. Rooth (1992) did, however, propose that the set of alternatives is a subset of the focus semantic value. This subset is restricted based on context. Some, including Rooth, have argued that these constraints are introduced to the discourse through presuppositions (Cohen, 1999; Rooth, 1992).

How and When are Alternatives Computed?

The process of how alternatives are computed is typically assumed to be semantically and/or pragmatically driven (Cohen, 1999; Fălăuş, 2013; Wagner, 2012). However, there have been also been theories that compute alternatives more on a syntactic level (Büring, 2016; Chierchia, 2006; Fox, 2007; Sauerland, 2012). It is clear that context plays an important role in deciding the alternatives, but to what extent the set of alternatives can be computed only on a semantic or even just syntactic basis is not clear.

As noted previously, it has been argued that alternatives are based on presuppositions (Cohen, 1999; Rooth, 1992). Cohen (1999) outlined a formal mechanism of how

alternatives come into play through presuppositions and how they can be computed compositionally. He argued that all elements in a set of alternatives share, what he called, a minimal presupposition. A minimal presupposition is a presupposition that does not entail anything else than what is entailed by the element for which we are computing the alternatives. The mechanism of determining the alternatives then relies on deciding the (minimal) presuppositions.

One area where there is substantial theoretical research on how alternatives are computed, is scalar implicatures. Formal theories on how scalar implicatures are computed can roughly be divided into three groups: pragmatic, lexical and grammatical theories (Sauerland, 2012).

Pragmatic theories essentially reason backwards and try to understand why a certain sentence is uttered (Geurts, 2010; Grice, 1989; Horn, 1972; Sauerland, 2004). They assume the speaker is rational and would have uttered the most informative, yet least costly alternative to the listener. This reasoning finds its origins in Gricean theory of conversation and often makes use of Grice's maxims of conversation, which are rather informally defined (Grice, 1975, 1989). These maxims state that an utterance should be informative, true, relevant and clear (Grice, 1989). In a first step, the set of potential alternatives of what the speaker could have said is computed. This set alternatives is then restricted through reasoning about why the speaker uttered a particular sentence. Under this header also fall game theoretic and Bayesian modeling approaches to scalar implicatures (Benz, 2010; Franke, 2011; Tessler et al., 2017). These approaches compute scalar implicatures through computationally modeling the interaction of rational speakers and listeners and their communicative goals in a restricted environment.

Horn (1972) was the first to define the notion of scales and his approach is also assumed under the pragmatic theories. His so-called *lexical scales*[2] consist of a limited number of lexical items that are ordered on an entailment relation or informativeness. Two of such scales that were defined by Horn (1972) are shown in (6) as examples.

(6) a. <all, most, many, some, few>

b. <and, or>

Both scales are ordered such that the left-most element is the most informative and thus the highest element on the scale. It entails all the elements lower on the scale. The elements on the same scale are called *scalemates*. (6a) defines a scale for quantifiers and (6b) a scale for the coordinating conjunction markers.

The computation of scalar implicature assumes the *competence assumption* (Horn, 1989; Sauerland, 2004; Van Rooij and Schulz, 2004). This assumption states that an alternative is changed from a weak implicature to a strong implicature by negating the alternative. It says that if a speaker does not believe an alternative to be true, the speaker must be sure the alternative is false. In other words, this assumes that if a speaker did not utter a specific alternative, this alternative must be false, because we would assume the speaker to only communicate what he or she believes.

[2] Also called Horn scales.

Consider (7)[3].

(7) a. Wilma read some of the papers

b. Wilma read all of the papers.

c. Wilma read most of the papers.

d. Wilma read many of the papers.

e. Wilma read few of the papers.

Assume that (7a) is uttered. Using the competence assumption we reason that the speaker does not believe that (7b), (7c) and (7d) are true, otherwise the speaker would have used that alternative. However, (7e) is already entailed by (7a) as it is lower on the lexical scale. Thus, we negate (7b), (7c) and (7d) and are left with the reading that Wilma read some but not many papers.[4]

It is generally assumed that computing pragmatic inferences is costly in terms of time, while semantic interpretations are less costly (Levinson, 2000). Lexical theories assume that scalar implicatures are default inferences and are stored in the lexicon Chierchia (2004); Levinson (2000). The scalar implicature, thus, is not computed on a global level, in contrast to pragmatic theories. In pragmatic theories scalar implicatures are always computed after processing the whole sentence.

An example of the computation of a scalar implicature at a non-global level is provided in (8)[5].

(8) If you take salad or dessert, you pay $20; but if you take both, there is a surcharge.

In this example the exclusive interpretation of *or* is computed before the end of the sentence, because the second part, where is explained what happens if one takes both, relies on this exclusivity of *or*. This example is thus evidence in favor of lexical theories which state that this exclusivity of *or* is the default inference and is thus computed on a non-global level.

Many experimental studies, however, support pragmatic theories on scalar implicatures and find that computing scalar implicatures is costly and takes longer to compute than the semantic interpretation of the sentence (Bott and Noveck, 2004; Degen and Tanenhaus, 2016; Huang and Snedeker, 2009). One hypothesis for this phenomenon is that the semantic meaning of a sentence is computed before its pragmatic meaning (Huang and Snedeker, 2009). For example, Degen and Tanenhaus (2016) conducted a visual world eye-tracking experiment on quantifiers, such as *all* and *some*, and also numbers. They found that the processing of scalar implicatures is affected by availability of alternatives from the earliest moment of processing. The availability of alternatives

[3]The example is taken from (Geurts, 2010).

[4]We also assume that Wilma did not read all papers, but this is entailed by the fact that she did not read many papers. This is actually a case of scale reversal, due to the negative polarity of the sentence. This means that the entailment relations between scalemates also reverses.

[5]The example is taken from (Chierchia et al., 2012).

influences the interpretation of the utterance and the processing time. They argue that scalar implicatures are computed in a constraint-based manner. That is, through constraints imposed by the context, the right alternative is chosen. In addition, processing speed is affected by the competition between alternatives and the probabilistic support in favor of them.

Finally, grammatical theories assume that scalar implicatures are computed using a covert grammatical operator *exh* (Chierchia, 2006; Fox, 2007; Sauerland, 2012). This so-called *exhaustivity operator* is defined as a conjunction of the proposition it applies to and a set of strong implicatures that follow from the assertion of the proposition in the context. They are derived in a similar fashion as we have seen in pragmatic theories. That is through reasoning about why the speaker uttered the particular proposition, making use of the competence assumption and/or Grice's maxims of conversation. An example of the semantic interpretation of the *exh* operator is shown in (9).

(9) $[[exh((7a))]] = [[(7a)]] \wedge [[\neg(7b)]] \wedge [[\neg(7c)]] \wedge [[\neg(7d)]] \wedge [[(7e)]]$

Here we use (7) as an example, where (7a) is the sentence that was uttered. (7b), (7c) and (7d) are derived strong implicatures and (7e) is entailed by (7a).

Whereas lexical theories cannot compute scalar implicatures on a global level and pragmatic theories cannot compute them on a local level, this *exh* operator allows for both. It can be applied to both constituents with a propositional meaning and sentences. The alternatives are thus computed on a more syntactic level and are separated from the reasoning about why speaker uttered a particular sentence.

2.1.2 Experimental Research

In this section, we consider experimental research that has looked at the general properties of alternative sets. We will start with looking at which kind of elements can be part of alternative sets. We then discuss how alternative sets are restricted. Finally, we will discuss how certain elements become more plausible in the set of alternatives.

Which Elements can be Alternatives?

Alternatives can be of a wide variety of syntactic categories and semantic types. Because of the constraints on our internal syntactic parser, all elements in a set of alternatives have to be of the same semantic type. For example, in research on focus, there are single elements that induce single element alternatives. This is called narrow focus. On the other hand propositions in focus can typically induce a much broader range of alternative propositions. This is called broad focus. Alternatives can thus also be whole propositions or sentences as we also have seen in Hamblin Semantics (Hamblin, 1973).

Buccola et al. (2018) argued that alternatives are not completely based on lexical properties, but more so on conceptual features and propose to use conceptual representations of words over the lexical realizations of the words. More specifically, they

claimed that alternatives differ from each other in primitive elements. These primitive elements are cognitively-inspired and denote basic natural classes such as *birds* or *berries*. But as they also noted, these claims are hard to verify as those conceptual properties are not visible directly on the words. The use of concepts therefore is more complex to measure in most experiments and especially for computational models that can only deal with the lexical realization of words.

Restrictions on Alternative Sets

A small body of experimental research has focused on characterizing the general restrictions on alternative sets. Most research so far has relied on a small set of fixed alternatives (Chemla and Singh, 2014). Only in recent years there have been a few attempts to characterize the general properties of alternative. The views on how alternatives sets are restricted can roughly be divided into a permissive view and a restrictive view (Katzir, 2013).

Rooth (1992) argued that there are no other restrictions than the ones that are contextually and pragmatically determined. In this permissive view the alternatives do not have to be contrastive. On the other hand Wagner (2005, 2006, 2012) noted that some alternatives are systematically excluded and argued that alternatives need to be mutually exclusive. In other words, alternatives need to be contrastive and thus need to be able to form partitions. Others have argued that non-contrastive alternatives can still be in the same set of alternatives Gotzner (2014); Gotzner et al. (2016); Husband and Ferreira (2016); Katzir (2013).

Gotzner (2014); Gotzner et al. (2016); Husband and Ferreira (2016) argued that there are two mechanisms behind the construction of alternatives. First there is activation of a broad set of alternatives, constructed using semantic, contextual and discourse information. This is then narrowed down by competition between the alternatives. More specifically this happens through inhibition of non-contrastive alternatives. They base these claims on online experiments where they induced focus using focus particles (Gotzner, 2014; Gotzner et al., 2016) and pitch differences (Husband and Ferreira, 2016).

The construction of the initial broad set of alternatives depends on contextual factors that determine the domain and quantification of the set of alternatives (Gotzner, 2014). As mentioned before, (Rooth, 1992) did not put any restrictions on this initial set. Several researchers have argued in favor of putting restrictions on this initial set. Gotzner (2014) argued that alternatives that are more readily available or salient in memory seem to be in the set of alternatives. Factors that have been shown to affect this availability in memory include how recent a possible alternative has been mentioned (Gotzner, 2014; Kim, 2012), contextual information (Byram-Washburn, 2013), discourse structure (Fraundorf et al., 2013; Kim, 2012) and world knowledge (Fraundorf et al., 2013; Kim, 2012).

Plausibility of Alternatives

Next to restrictions on the set of alternatives, the set itself is normally ordered on plausibility. The factors influencing the plausibility have not been studied systematically or formalized. Kruszewski et al. (2016) were, to our knowledge, the first that studied the plausibility in a more systematic way, by computationally assigning plausibility ratings to alternatives and ordering the alternatives accordingly. Note that the ordering of the set of alternatives in this case is different from the orderings as defined by some kind of scale for scalar implicatures. In terms of plausibility within a set, alternatives are more salient than others. We have already mentioned that alternatives that are more readily available in memory through recent mentioning, world knowledge and discourse structure, tend to be more plausible.

Adults, in contrast to children, seem to have a preference for upper bound interpretations of scalar implicatures (Noveck, 2001). Thus, alternatives that define some upper bound are preferred over alternatives that define a lower bound. This distinction can be made clear using the widely studied quantifier *some*. Adults tend to interpret this quantifier as *some, but not all*, thus defining an upper bound. Children, on the other hand, tend to interpret this quantifier as *at least one*. This distinction is attributed to the fact that adults have more pragmatic reasoning tools available and children rely more on the logical interpretation, rather than the pragmatic interpretation (Noveck, 2001).

2.1.3 Statistical and Computational Models

The distributional semantic (DS) framework offers a way to capture similarity between words and could therefore be very suitable for modeling alternatives. This is because the similarity measures in DS capture the gradedness in plausibility between alternatives and also because the notion of similarity in DS measures how similar the contexts of two concepts are in which they occur. Alternatives to a concept are known to occur in similar contexts as the concept itself. By approximating over many examples, the pragmatic and contextual tendencies are filtered out and the general semantic tendencies of a word are approximated. Thus, DS offers us a way to study alternatives in terms of their semantic and thus contextually-invariant properties.

Kruszewski et al. (2016) introduced the possibility to use DS to computationally model alternatives. They focused on alternatives that are induced by negation and found DS particularly suitable for modeling alternatives. Their study focused on negating common nouns. Follow-up studies have focused on negating adjectives (Aina, 2017) and negating verbs (Storozum, 2018). These studies have also shown that alternatives can be modeled using DS. Moreover, Aina (2017) showed that the scalar properties of alternative sets can be captured using DS and she has reconstructed several adjectival scales. All these approaches have focused on negation as alternative-inducing element.

What these studies tell us, is that alternatives can be approached probabilistically and that alternatives are close to each other in the semantic space. Thus by considering words in context, we are able to extract relevant alternatives. This view is not just

restricted to DS models and could potentially be extended to other types of LMs which build representations by considering words in context. In addition, the probabilistic approach of alternatives allows us to order them on their likelihood in a given context. Potentially we are thus able to capture scalar effects of certain elements. As discussed, this has indeed been captured by Aina (2017).

Another use of computational models in research on alternatives has been to validate or invalidate experimental hypotheses. We describe here a pragmatic approach to alternatives and how this is modeled computationally. Recall that pragmatic theories reason backwards about why a speaker uttered a particular utterance and this can be explained with Gricean maxims (Grice, 1989). This pragmatic view has been modeled within several specific and small contexts with Bayesian statistics in the Rational Speech Act (RSA) framework (Frank and Goodman, 2012). This approach has its foundations in game theory. In the RSA framework a speaker uses probabilistic reasoning to infer the state of the listener, while the listener creates a model of the speaker's internal state. The speaker then uses this model of the listener to convey the message to the listener in the most communicative and efficient way. The listener tries to reason about what the speaker could have meant. Bayesian methods are especially suitable for modeling these constraints on communicating a message (Franke and Jäger, 2016).

Tessler et al. (2017) looked at the process of how an alternative becomes salient in context and viewed this process as a pragmatic inference. Participants were given a statement, containing a scalar adjective, (e.g., *It's warm.*). Next, they were given two sentences and asked to judge which sentence rephrases best what the statement could have meant, given a one-sentence context (e.g., *Tanya lives in Maryland and steps outside in Winter*). The choices were always a subordinate comparison class (e.g., *It's warm, relative to other days in winter*) and a superordinate comparison class (e.g., *It's warm, relative to other days of the year*), relative to the context sentence. Their Bayesian model, that is an extension of the RSA model for gradable adjectives (Lassiter and Goodman, 2013), was able to explain the human judgments. They concluded from the behavior of their model that listeners combine both category knowledge and contextual knowledge to decide which alternatives are plausible and become salient. The study by Tessler et al. (2017) thus used a computational model to study factors that influence the set of alternatives.

There have been a few other studies that use game theoretic approaches (Benz, 2010) and Bayesian modeling approaches (Franke, 2011; Frank and Goodman, 2012; Goodman and Stuhlmüller, 2013) to model scalar implicatures. These approaches will not be discussed here.

2.2 Focus Particles

Focus particles are words, such as *even*, *only*, *also*, *too*, *just* and *merely*. They are used to put focus on a particular element and contrast its properties to elements in a set of alternatives. Focus is traditionally seen as a grammatical category that can influence

the interpretation of a sentence, depending on the place of the focus. It can be used to disambiguate a sentence or change the current interpretation.

This section will provide an overview of general syntactic, semantic and pragmatic properties of focus particles in English and will also focus on two particular particles: *even* and *only*. Note that the cross-linguistic behavior of at least some of the focus particles might differ from what is described here. Also note that focus particles are not the only way to put an element in focus and contrast it to a set of alternatives. Beaver and Clark (2008) have given an overview of other expressions that are sensitive to focus, which include negative particles, generic sentences, quantificational adverbs and determiners like *many* and *most*.

2.2.1 Syntax

Focus particles are most often seen as a subclass of adverbs, but some of their syntactic properties make them a distinct subclass within adverbs (Quirk et al., 1985). Focus particles can modify any syntactic phrase and are thus very flexible in the position in which they occur in the sentence. A *constituent focus particle* modifies a single constituent and a *sentential focus particle* modifies a sentence or clause. Both *even* and *only* have both sentential and constituent variants. In (10)[6], we can place the particle *only* at every position in the sentence.

(10) She told him that she loved him.

Focus particles thus mark a particular constituent or phrase. As we noted before, in English, focus is induced through placing prosodic prominence on the constituent or phrase in focus. (12) is the underlying proposition that *only* modifies in (11) and is called the *prejacent*.

(11) John loves only $[\text{Mary}]_F$.

(12) John loves Mary.

Finally, focus particles can only associate with constituents that they c-command (Bayer, 1996; Jackendoff, 1972; Rooth, 1985). This means that a focus particle can only associate with its sisters and their descendants in the syntactic tree of the sentence. This also means that focus particles can be ambiguous in which constituents they associate with and disambiguation happens through marking of one of the constituents that the focus particle can associate with. Consider (13), where (13a) is ambiguous between (13b) and (13c).

(13) a. John only dislikes cats.

b. John only $[\text{dislikes}]_F$ cats.

c. John only dislikes $[\text{cats}]_F$.

[6]The example is taken from (Erlewine, 2018).

In this sentence, the focus particle is in verb position. Both the verb and the object are c-commanded by the focus particle. (13b) falsifies the expectation that John's attitude toward cats is worse than just disliking. He was expected to have an absolute hatred toward them. (13c) on the other hand suggests that John does like other animals than cats.

There are, however, exceptions to this c-command constraint. Jackendoff (1972) noted that *only*, in contrast to for example *even*, cannot modify a leftward subject. In (14)[7,8], *even* can modify the leftward subject *a professor*, while *only* cannot.

(14) a. A [professor]$_F$ seems to even be at the party.

b. #A [professor]$_F$ seems to only be at the party.

Erlewine (2014) has proposed that *even* can associate with elements that are outside its scope, if the original position (before movement) was within the scope of *even*. *Only* does not have this property.

There is some research on the influence of the syntactic position of the focus particle on the comprehension of the whole sentence. Kim (2011) found that the children of age 4 and 5 had no problems in computing the meaning of *only*, but had trouble in assigning the right scope of *only*. The children tended to prefer object scope, even when *only* quantified over an element in subject position. There is some evidence that the syntactic position of the focus particle influences the sentence comprehension in adults as well (Paterson et al., 2007).

2.2.2 Semantics and Pragmatics

There has been a lot of debate on the exact meaning contribution of focus particles. Because of the variety of focus particles and their complexity, this is best discussed on a case-by-case basis. Here, we will describe some general properties focus particles tend to have.

The first thing to note, is that the only contribution of focus particles is their operation on the set of alternatives. The effect of adding a focus particle to a sentence cannot falsify the prejacent, since the prejacent will always hold. Consider the sentences in (15).

(15) a. John likes horror movies.

b. John even likes horror movies.

c. John also likes horror movies.

d. John likes horror movies too.

e. John only likes horror movies.

[7]The example is taken from (Erlewine, 2014).

[8]Ungrammatical and unacceptable sentences are marked by a "#" at the start of the sentence.

f. John just likes horror movies.

g. John merely likes horror movies.

Here (15a) is the prejacent. This prejacent is still true when we add any focus particle. It cannot be canceled out, for example, by adding the negation *but he doesn't* to the sentences. What is instead modified, is the set alternatives of *liking horror movies*.

Focus particles can be *inclusive*[9] or *exclusive*. Inclusive particles are used to state that the properties of the element in focus are shared by the set of alternatives. Exclusive particles have the opposite effect. They state that there is no alternative that shares the same properties. *Even*, *also* and *too* are generally considered to be inclusive focus particles, and *only* and *just* to be exclusive focus particles. Focus particles can also be scalar. In this case the alternatives are ordered on some kind of scale. We will discuss the research on scales in focus particles in more detail when discussing the focus particle *even*.

The general consensus is, that focus particles introduce linguistic expectations in the form of presuppositions. These expectations can influence the acceptability of sentences and depend on the alternatives that are considered in the given context. As noted before, there is generally very little research on which alternatives are considered. This is not much different for the set of alternatives, induced by focus particles. As we have discussed before in the framework of Alternative Semantics, Rooth (1992) stated that the set of alternatives is restricted by context and thus pragmatically determined. An alternative was provided by Beaver and Clark (2008), who argued that the set of alternatives is always a *Question Under Discussion* (Roberts, 1996). Under this view the set of alternatives is decided by the interlocutors through communication and by establishing a common ground, and is seen as a little language game. The goal of the interlocutors is to restrict the set of alternatives by choosing a strategy and subsequently asking questions.

Filik et al. (2009) studied online sentence comprehension of sentences with focus particles *even* and *only*. They did an eye-tracking study and tracked the eye movements of participants, while they were reading sentences containing *even* and *only*, and where the element in focus was congruent or incongruent with the context. They used the observation that incongruent information requires more rereading and takes longer for participants to read. They found that the semantic interpretation of both focus particles was computed online, but that the computation of *even* was delayed, while the computation of *only* was computed rapidly. These results suggest that not all focus particles are computed in the same way and also that the semantic interpretation of *even* is more complex than the semantic interpretation of *only*, possibly due to the element of surprise in *even*.

[9] Inclusive particles are sometimes also called *additive* particles.

2.2.3 Computational Analyses

There are only a few studies analyzing the behavior of focus particles from a computational perspective. One of the few computational studies on focus particles focuses on the scalarity of *even* (Gast and Rzymski, 2015). The authors investigated the contextual factors influencing the scale of *even*. In particular, they looked at the influence of the words around *even* on an attitudinal scale of *even*. Their attitudinal scale is defined as "even better" or "even worse". By computing sentiment values for the context words and correlating them with their attitudinal scale, they drew conclusions on whether context influences this attitudinal scale and thus that context influences the scale of *even*. They found a significant correlation between the positive and negative sentiment scores and the sentence-level propositions which define their attitudinal scale, meaning that context influences the scale of *even*.

2.2.4 Relation between *even* and *only*

Before moving in detail into the focus particles *even* and *only*, we want to point out that *even* and *only* are often seen as being opposites in English. Besides these particles being the most-studied focus particles in English, this was also motivation to focus on these two particles. Horn (1969) noted there is a parallelism between *even* and *only* in what they presuppose and assert. Specifically Horn claims that "*even* asserts what *only* presupposes and presupposes the negation of what *only* asserts" (Horn, 1969, p. 106). Beaver and Clark (2008) noted that *even* and *only* have significantly different semantic properties and therefore called them pragmatic antonyms. Zeevat (2009, 2013) summarized the behavior of *even* and *only* using mirativity as in "more than expected" and "less than expected", respectively.

Greenberg (2018, 2019) also noticed the parallelism between *even* and *only*, but argued that the relation between *even* and *only* is more complex. She argued that even though *even* and *only* contrast each other in the scalar ordering, they do not contrast each other in terms of mirativity, because *only* cannot be fully explained in terms of mirativity (Greenberg, 2019; Roberts, 2011). *Even* and *only* also do not contrast each other in terms inclusivity, because it has been argued that *even* is not fully inclusive (Greenberg, 2016, 2019; Krifka, 1992; Rullmann, 1997, 2007; Von Stechow, 1991). Greenberg (2018, 2019) furthermore used these observations in arguing for a hybrid semantic interpretation for *only* that combines the scalar semantic interpretation with the non-scalar semantic interpretation.

2.2.5 *Even*

Informally, *even* denotes that the element that follows it is less likely or more surprising than we would expect, given our knowledge of the world. The exact meaning contribution of *even* is not agreed upon in the literature. Most do agree that the prejacent of *even* is asserted. What is furthermore generally agreed upon, is that *even* compares some property on some kind of scale and is the least plausible or least expected element on this scale. Also, *even* has an inclusive meaning.

On a syntactic level, there is debate whether sentential *even* can take surface scope. Kay (1990) said that *even* always takes wide scope, due to *even* not having any semantic contribution. The pragmatic contribution requires it to take wide scope. Rullmann (2007) said that alternatives of *even* are computed at the constituent level, but that scalar inferences can be made at sentence level. Erlewine (2016) on the other hand argued that *even* in VP position always takes surface scope.

Traditionally, formalizations of *even* have taken a form similar to (16) (Karttunen and Peters, 1979; Rooth, 1985, 1992).

(16) $[[\text{even } p(x)]] =$
Existential presupposition: $\exists q\,[q \in A \wedge q \neq p \wedge q(x)]$
Scalar presupposition: $\forall q\,[[q \in A \wedge q \neq p] \rightarrow q(x)]$
Assertion: $p(x)$

Note, however, that there has been a lot of debate on what kind of presuppositions *even* introduces and, moreover, whether *even* introduces any at all. Karttunen and Peters (1979) were the first to argue that *even* introduces two different presuppositions. This has also been the traditional and dominant viewpoint. Firstly, *even* introduces an existential presupposition, which says there is at least one other alternative in the set of relevant alternatives A, which shares the properties over which *even* quantifies. Thus, this defines the inclusive meaning, usually assigned to *even*. The second presupposition is a scalar presupposition that states that the property that is applied to x is the least plausible alternative. All other alternatives in a particular context (A) are stronger.

The inclusive meaning of *even* has been challenged by many (Greenberg, 2016, 2019; Krifka, 1992; Rullmann, 1997, 2007; Szabolcsi, 2017; Von Stechow, 1991; Wagner, 2014). For example, consider the little discourse in (17)[10]. The properties of being an assistant professor and associate professor are mutually exclusive and thus cannot be both true. Therefore, *even* cannot have an inclusive meaning in this particular example. Interestingly there are also examples where *even* can be used, while *also* and *too* cannot be used, such as in (18)[11].

(17) A: Is Claire an [assistant]$_F$ professor?
B: No, she is even an [associate]$_F$ professor.

(18) a. He is not (just) an assistant professor, he is even an associate professor.

b. #He is not (just) an assistant professor, he is also an associate professor.

Rullmann (2007) argued that the existential presupposition can be derived from the scalar presupposition and should therefore not be lexically specified as a presupposition, as in (16). Wagner (2014, 2015) noted that the behavior of *even* is different, depending on the syntactic position. There is an existential presupposition when *even*

[10]The example is taken from (Rullmann, 1997).
[11]The example is taken from (Rullmann, 2007).

modifies a determiner phrase (DP), but there is not always one when it modifies a verb phrase (VP). Szabolcsi (2017) argued that the existential presupposition comes into play through the activation of alternatives through focus. One effect of the meaning contribution of *even* is that at least one other alternative in the set of focus alternatives, other than the prejacent, must be true. The reason *even* is not always inclusive, is that it does not always directly modify the element it associates with. To make this more clear, consider (19)[12] and consider the fact that Eeyore normally loves eating thistles.

(19) Those thistles must be really prickly! Even Eeyore spit them out!

In principle *Eeyore* could have been the only one eating *those thistles*. The property that *even* modifies, is the edibility of the thistles.

The element in focus is often the lowest element on some kind of scale. This is seen by most as a presupposition (Horn, 1969; Karttunen and Peters, 1979; Rullmann, 1997), like we have characterized it before in (16) as well. However, there are many cases where it is arguably not the lowest element on the scale. For example, consider the sentence in (20).

(20) Bill even won bronze.

It should be clear that less plausible alternatives are Bill winning silver or gold. For this reason, Kay (1990) has argued that the set of alternatives is restricted to only those contextually relevant or salient. In this example, we could argue that we did not expect Bill to win even a medal and therefore a gold or silver medal are not salient alternatives. We have also already discussed another potential method to solve this. As mentioned before, Beaver and Clark (2008) stated that the set of alternatives are a QUD and it is restricted through interaction. Finally, we could reformulate the formula of the scalar presupposition, so that the scalar presupposition does not quantify over all alternatives anymore. In other words, the universal quantifier should be replaced. This option has not been preferred in the literature (see e.g., Greenberg, 2019).

What kind of scale the alternatives of *even* are ordered on and how this scale is ordered are not very clear either. The traditional assumption is that it is some kind of likelihood scale (Chierchia, 2013; Karttunen and Peters, 1979; Rooth, 1985, 1992). Fillmore (1965) argued that *even* is used to indicate a violation of some kind of expectation and thus that the scale is ordered on unexpectedness of the focus element being true. Fauconnier (1976) said that the scale is contextually determined and that the alternatives are ordered on a pragmatic entailment scale. Kay (1990) ranked the alternatives on informativeness. Herburger (2000) claimed the scale of *even* is ordered based on noteworthiness. Finally, various researchers have argued that the scale of *even* is based on a contextually salient and graded property (Greenberg, 2015; Rullmann, 2007).

All these claims have not seen a proper formal treatment and it could be said that some claims share similarities in their definition of the scale. Greenberg (2015) tried to give a revised formula for the scalar presupposition. One which bases the scale on

[12]The example is taken from (Szabolcsi, 2017).

a contextually salient and graded property, but it is still not clear how we derive the contextually salient property and how alternatives are judged on this.

The scale of *even* reverses in downward-entailing or negative environments (Karttunen and Peters, 1979). Two explanations have been offered in the literature. It has been argued that *even* is lexically ambiguous (Giannakidou, 2007; Herburger, 2000; Rooth, 1985; Rullmann, 1997). *Even* can be a negative polarity item (NPI) as well as a positive polarity item (PPI). This difference is lexicalized in other languages such as Dutch (Rullmann and Hoeksema, 1997), Greek (Giannakidou, 2007) and Spanish (Lahiri, 2008). We will mention Greek in particular where there are three different lexical items for *even*, which roughly correspond to the PPI version of *even*, the NPI version of *even* and a version of *even* that allows the use of a more flexible scale than the standard traditional likelihood scale (Giannakidou, 2007). This flexible scale is contextually specified and salient in the context.

Another stance has suggested that the scale reversal of *even* is due to *even* taking scope over the particle or constituent that caused the downward-entailing (DE) environment (Karttunen and Peters, 1979; Wilkinson, 1996). As we have mentioned previously, there is debate on what scope sentential *even* can take. This argument relies on the fact that *even* can take wide scope and thus can take scope over the "negative" particle. The fact that *even* can take wide scope thus causes *even* to be ambiguous in having a scale that can be ordered in two ways.

Finally, Rullmann (2007) argued that neither of those two theories are correct and that *even* causes neither lexical, nor scope ambiguity. He instead argued that the end of the scale is lexically unspecified and thus determined by the context. In upward-entailing (UE) or positive environments, due to the absence of an element that could reverse the scale and the presence of only "positive" elements, there is only one way to order the scale. In DE environments, the "negative" element would cause the scale to reverse.

2.2.6 *Only*

Informally, *only* has an exclusive interpretation where all relevant alternatives, but the one *only* associates with, do not have a certain property. Like with *even*, there is still a lot of debate on what the exact meaning contribution of *only* is. In particular there has been a lot of debate on whether *only* has scalar properties and to what extent they need to be encoded in the semantics of *only*. Consider (21) where *only* does have a scalar interpretation.[13] On the other hand, in (22) *only* does not have a scalar interpretation.[14]

(21) At the time of the battle of Shiloh, MacPherson had only been a lieutenant colonel.

(22) Because of the poor state of the roads, most of them can only be used during the dry season.

[13]The example is taken from (Beaver and Clark, 2008).
[14]The example is taken from the web.

On a syntactic level, sentential *only* always takes surface scope (Rooth, 1985; Taglicht, 1984). Constituent *only* in non-subject position can introduce scope ambiguities. Bayer (1996) noted a difference between subjects of finite and nonfinite clauses. When *only* modifies the subject of a finite clause, it cannot take scope over elements outside the clause, which is possible in nonfinite clauses.

Klinedinst (2005) noticed that the prejacent of *only* together with the exclusiveness property cannot account for all data. For example, in (23)[15], the fact that John only has a master's degree does not exclude that he also has a bachelor's degree.

(23) a. John only has a master's degree.

b. #John does not have a bachelor's degree.

Instead, *only* seems to only exclude alternatives that are higher on a certain scale. Klinedinst (2005) concluded from this that *only* also has a scalar interpretation. Beaver and Clark (2008) also noted that *only* can have both scalar and non-scalar uses and asked the question, in a similar fashion as with *even*, whether there are multiple lexical entries for *only* and thus whether *only* is ambiguous between a scalar and a non-scalar interpretation.

Traditionally, there have been two lexical entries for *only*: a non-scalar one as in (24) and a scalar one as in (25).

(24) $[[\text{only}_{ns} p(x)]] =$
Presupposition: $p(x)$
Assertion: $\neg\exists q[q \in A \wedge q \neq p \wedge q(x)]$

(25) $[[\text{only}_{s} p(x)]] =$
Presupposition: $p(x)$
Assertion: $\neg\exists q[q \in A \wedge q \neq p \wedge q >_s p \wedge q(x)]$

(24) says that there is no other contextually relevant alternative, defined by the set A that applies to x than p. (25) says that there is no property in the set of relevant alternatives A that is more plausible than p that is applied to x. Finally, both versions presuppose the prejacent $p(x)$. This presupposition has been subject to debate.

Next to the prejacent of *only* being a presupposition (Horn, 1969; Rooth, 1985, 1992), it has been argued that the meaning contribution of the prejacent of *only* is a weaker existential presupposition (Horn, 1996; Zeevat, 2009, 2013), a conversational implicature (McCawley, 1981; Van Rooij and Schulz, 2007), a projection[16] (Roberts, 2011), an entailment (Atlas, 1993) and a mirative implication[17] (Beaver and Clark, 2008).

[15] The example is taken from (Winterstein, 2012).

[16] A presupposition that cannot be canceled out by other clauses is said to be a projection. This is thus a slightly stronger version of a presupposition. However, contrary to a presupposition, a projection does not necessarily have to be information that part of the speaker's beliefs or background knowledge. It could also be new information.

[17] The prejacent is less than expected.

Horn (1969) has argued that *only* has a non-scalar interpretation in subject position and a scalar interpretation in non-subject positions. Many (e.g., Beaver and Clark, 2008; Coppock and Beaver, 2014; Roberts, 2011) have argued for only a scalar version of *only*. This scalar version is then used to explain the non-scalar examples. The alternatives in examples are said to still lie on a scale, more specifically, an entailment scale.

Coppock and Beaver (2014); Beaver and Clark (2008) used the notion of Question Under Discussion (QUD) to define the lexical entry of *only*. They argued that the meaning contribution of *only* can be defined by an *at least* component which is presupposed and an *at most* component which is asserted. The *at least* contribution states that there an answer to the QUD that is at least as strong (on a particular type of scale) as the prejacent. The *at most* contribution states that no answer to the QUD is stronger (on a particular type of scale) than the prejacent.

Other approaches to *only* did not use any notion of scales (e.g., Krifka, 1992; Rooth, 1985, 1992; Winterstein, 2012; Zeevat, 2009, 2013). Zeevat (2009, 2013) regarded both *even* and *only* as *mirative markers*. Mirative markers mark surprising or unexpected elements. In case of *only*, Zeevat (2009) said that the prejacent of *only* is less than expected. An expectation is weakly presupposed and the falsehood of this expectation is asserted. A weak presupposition does not have to be common ground or part of the speaker's beliefs. The mirative effect of *only* then comes from the interplay between the weak presupposition and the assertion. There is a certain expectation that is presupposed and that is denied by using *only*.

Winterstein (2012) argued that the scalarity of *only* is not part of the semantics but arises as part of the pragmatic interpretation. He argued for the same semantic representation as Zeevat (2009, 2013). The scalar effects of *only* are the result of the pragmatic effects of the exclusiveness of *only* and the contrastiveness of the alternatives. Reconsider (23a), which has been repeated in (26).

(26) John only has a master's degree.

This sentence does not exclude that John has a bachelor's degree, because having a bachelor's degree is not distinctive from having a master's degree. That is the case because having a master's degree implies that one also has a bachelor's degree.

Greenberg (2018, 2019) proposed a lexical entry for *only* that is somewhat in the middle between a scalar and non-scalar entry. Her proposal is based on an earlier version of *only* from Guerzoni (2003). The formula is outlined in (27).

(27) $[[\text{only}_s p(x)]] =$
Presupposition: $p(x) \wedge \forall q[[q \in A \wedge q \neq p] \rightarrow q >_s p]$
Assertion: $\neg\exists q[q \in A \wedge q \neq p \wedge q(x)]$

(27) presupposes the prejacent, although this could be omitted according to Greenberg (2018, 2019). More importantly, it also presupposes that all contextually relevant alternatives of some property, defined by the set A, are higher on some kind of scale. Thus, stating that the this property is "weaker" than all relevant alternatives. It furthermore

asserts that there is no such "stronger" property that holds for the associate of *only*. Greenberg (2018, 2019) furthermore made the comparison of this *hybrid* interpretation to the semantic meaning of *even*. The universal quantifier in her formula does not scope over all alternatives, but only over all contextually relevant or salient alternatives. Thus, like Kay (1990) argued for *even*, she argued in a similar fashion that the set of alternatives is restricted by the context and the prejacent does not necessarily have to be the lowest element on the scale. But it should become the lowest element when considering the right context.

2.3 Language Models

A language model (LM) is a model that is able to learn syntactic, semantic and possibly pragmatic patterns from natural language data. In its pure form, an LM is a probability distribution over all units in the language, usually sequences of words or characters.[18] In this section, we will only consider word-level LMs.

LMs have been used in a wide variety of applications, ranging from machine translation and speech recognition to information retrieval. Since they have been trained on human language, they contain some sense of how language is structured and used. LMs have been very successful in part-of-speech tagging (Ling et al., 2015b; Plank et al., 2016), parsing (Charniak et al., 2016; Fried et al., 2017), machine translation (Luong et al., 2015; Schwenk et al., 2012), speech processing (Arisoy et al., 2012; Mikolov et al., 2010), question answering (Rajpurkar et al., 2016) and other tasks where the syntactic structure of a sentence is predicted.

The more traditional view on LMs, is that they predict a probability distribution over words, given a sequence of words. Traditional LMs use count-based approaches, such as n-grams, to induce this probability distribution. In n-grams, this probability distribution is only based on the previous n words.

Neural LMs rely on the distributional hypothesis, which states that the semantic meaning of a word can be represented by its context and thus that semantically similar words occur in similar contexts. Through neural networks, the neural LMs induce a continuous probability distribution over words.

Word embeddings are the most standard version of this type of LMs. They represent words as a high dimensional vectors. These vectors are trained on natural language data using neural networks. Word2vec is the primary and most-used word embedding model (Mikolov et al., 2013). Because word order is less taken into account in most of these models, they are, in essence, more semantically oriented. They have, however, also been shown to be capable of representing syntactic knowledge with a few simple modifications (Ling et al., 2015a).

Recurrent neural networks (RNNs) are a type of neural networks where nodes get input from previous nodes. In an RNN LM, this means that, at the prediction of each word, the model has information about the previously seen or generated words in an

[18]There are many other uses of LMs that are based on, for example, sequences of sentences, paragraphs or classes

utterance. In more technical terms, the RNN has sequential input. This means, that the input is separated into time steps. It encodes each step in terms of a vector. The information is then passed on to the actual model that is trained on gold-standard data. The output is then decoded to a human-readable format, which, in a neural LM, are words or any other linguistic structure.

A major problem for the standard RNNs is known as the vanishing gradient. As a consequence of many computations, previously computed information gets filtered out. This results in the fact that RNNs are not able to learn very long-distance dependencies. To overcome this, long short-term memories (LSTMs) have been developed (Hochreiter and Schmidhuber, 1997). An LSTM is a special version of an RNN, that adds an extra component that regulates what and how much information should be remembered and forgotten, and is thus able to selectively add and forget information. Through this, it is capable of capturing longer-distance dependencies. Bidirectional versions of the model do not only process a sentence forwardly, but in addition also backwardly (from the end to the start of the sentence) and combine both sources of information. This way, the predictions of the model are based on more information and also are not incremental anymore.

Since humans process sentences mostly incrementally (on a word-by-word basis) (Kamide et al., 2003; Tanenhaus and Brown-Schmidt, 2008; Tanenhaus et al., 1995; Traxler et al., 1997), this type of models match the human sentence processing especially well and they have therefore been very successful in modeling language. RNNs and LSTMs have been shown to be able to catch syntactic patterns and learn grammaticality judgments (Bernardy and Lappin, 2017; Chowdhury and Zamparelli, 2018; Dyer et al., 2016; Gulordava et al., 2018; Linzen et al., 2016).

Lau et al. (2016) argued that linguistic knowledge can be probabilistic and show a high correlation between grammaticality judgments by their LM and human judgments. Chowdhury and Zamparelli (2018) used a simple LSTM and showed that this LM is able to separate grammmatical sentences from minimally different ungrammatical sentences to some extent, on three different syntactic phenomena: subject versus relative clause processing, wh-gaps and syntactic islands. Whereas traditional count-based LMs have difficulties in capturing long-distance dependencies, neural LMs have been shown to be more capable of capturing these (Linzen et al., 2016; Wilcox et al., 2018).

Chapter 3

Experimental Setup

The experimental setup is divided into two parts. In the first part we investigate the differences in alternative sets that are induced by the focus particles *even* and *only* and explore to what extent the computational models are able to capture these differences. We will refer to this part as *Predicting Focus Particles*. In the second part we investigate the scalar effects of *even* on the set of alternatives that the model generates. We will label this part as *Generating Alternatives*.

We start by explaining the reasoning behind the setup of the experiments. Next, we explain how we acquired the data, which we used in collecting human judgments. The human judgment tasks will be explained thereafter. We then move on to explain which computational models were used and how we evaluated the performance of these models. Finally, we explain how we extract alternatives from the models.

3.1 Motivation and Overview

3.1.1 Predicting Focus Particles

As we have seen, focus particles mainly operate on the alternatives that are considered. We have also seen that focus particles can be ambiguous in the element they associate with and that intonational cues that induce focus disambiguate the sentence. This means that, in order to use focus particles as part of our setup, we have to use them in contexts that are fairly unambiguous, because in written text there is no way to disambiguate a sentence using focus marking.

The elements in a set of alternatives can be varied, but focus particles can severely restrict this set. Because the meaning contribution of focus particles is to put restrictions on the set of alternatives and they can be used in similar contexts, we can create a setup where we have a fixed context and only alter the focus particle. We can directly compare the differences between the focus particles and with that attribute any differences in the acceptability of a sentence to the different operations on the alternative sets. To maximize this difference, we use two focus particles that seem to have almost opposite behavior: *even* and *only* (Beaver and Clark, 2008; Greenberg, 2018, 2019;

Horn, 1969; Zeevat, 2009, 2013). Both the inclusive *even* and the exclusive *only* also occur frequently enough in language for the computational language models (LMs) to build a reliable representation of these words.

Focus particles also show very flexible syntactic behavior. This allows us to study their effects on different syntactic positions, as alternative sets might be influenced by the syntactic position. For example, *only* in subject position could have a scalar interpretation, while this is less likely when *only* occurs in object position. Another example is that a focus particle that modifies a verb, will induce verbal alternatives, while a focus particle in subject or object position will modify nouns or other elements (such as adjectives) that modify the noun.

We therefore study three different syntactic positions where a focus particle could occur: in subject, verb and object position. The structure of the sentences is shown in (28).

(28) a. Even/Only *Subject Verb Object*.

b. *Subject* even/only *Verb Object*.

c. *Subject Verb* even/only *Object*.

Note, that a focus particle in verb position could also modify the object and is thus ambiguous between the two readings. We therefore systematically collect sentences with the focus particle in either of the three positions. The collection of the data will be further discussed in Section 3.2.1. We then remove the focus particle and give these sentences to human participants to judge which focus particle between *even* and *only* is the most plausible in the sentence.

Next, we evaluate the computational models on the sentences with both the correct and incorrect focus particles as judged by the participants. We split this analysis into three parts. First, we compare how predictable the sentences are with the correct and incorrect focus particle, as judged by our participants. Throughout this thesis, we will refer to this as *naturalness*. We would expect that sentences with the correct focus particle are more predictable for the models and thus perceived as more natural by the models. Of course this is not always the case as sentences can be ambiguous and both focus particles might be equally plausible in some contexts. For this reason, we remove sentences that did not have full agreement on the judgments of the focus particle from the analysis.

In a second experiment, we isolate the exact meaning contribution of the focus particle in a particular sentence and compare the correct and incorrect use of the same focus particle. Here, we would expect that using the correct focus particle has a more positive effect on the naturalness of the sentence than using the incorrect focus particle.

Finally, we investigate how well the models are able to predict the correct focus particle. We measure the accuracy of how often the human judgments correspond with the models' judgments. The models make this decision, based on the most natural sentence between the sentences with *even* and *only*. Because the only difference is in the focus particle, we can attribute any differences in naturalness of the whole sentence

to the difference in focus particles. If the judgments of the models largely correspond with the judgments of the participants, we can conclude that the models are able to capture the same intuitions on focus particles as humans do. Using our reasoning from before, we can then also conclude that the models are able to learn some human intuitions about alternatives.

3.1.2 Generating Alternatives

We are interested in the extent to which computational models are able to generate alternatives that are plausible from a cognitive and linguistic point of view. One way to do this, is to find ways to have human participants generate alternatives and have the computational models generate alternatives, and find a way to compare the two.

Again, the range of plausible alternatives is very large. Therefore, we need to come up with a way to restrict the possibilities. By explicitly providing an alternative that is in the set of alternatives, we restrict the range of potential alternatives. An example is provided in (29).

(29) John likes beer and ___.

The set of alternatives is not only restricted to things John likes, but is also restricted to things John likes to drink.

By using the focus particle *even*, we are able to restrict this set of alternatives further. Consider the example in (30).

(30) John likes beer and even ___.

The set of alternatives is not only restricted to things John likes to drink, but in addition to things we think are less likely for him to drink than beer. One option could be a stronger alcoholic drink, such as *absinthe*. The alternative in this case is less likely or less expected than the first alternative that is provided. *Even* also creates an asymmetry between the alternatives, since they are ordered on a scale. Switching the order of the two alternatives in (31), as in (32), would make the sentence less natural.

(31) John likes beer and even absinthe.

(32) John likes absinthe and even beer.

We could still think of a situation where John has an expensive taste and would normally only drink expensive alcoholic drinks. In this case the scale changes from the strength of the alcoholic drink to the expensiveness of the alcoholic drinks. In our current setup, we cannot fully control or prevent this change of scales from occurring, because we do not know which alternatives the participants will generate.

There are other examples where the scale change of *even* is more apparent, such as the sentence in (33).

(33) The expert technician services computers and even televisions.

Switching *computers* and *televisions* would make the sentence perfectly acceptable as it would only change what we would expect to be the expertise of the technician.

In general, we might still see an effect of changing these two elements, which would still be an indication that the models learn something about the plausibility of alternatives in a certain context. This brings us to our first experiment where we compare the naturalness of a sentence in normal order to the sentence where the order of the alternatives is switched. This setup is shown in (34).

(34) a. Condition 1: *Subject Verb Object* and even *Alternative*.

b. Condition 2: *Subject Verb Alternative* and even *Object*.

In a similar experiment, we leave out either the item or the alternative and compare the acceptability of those two sentences. This setup is shown in (35).

(35) a. Condition 1: *Subject Verb Object*.

b. Condition 2: *Subject Verb Alternative*.

Since the generated alternative should be the less expected element, we would predict that the sentence with this alternative would be more surprising, and should therefore be less predictable andthus perceived as less natural by the LMs. An example of this is given in (36).

(36) a. Australians eat beef and even crocodiles.

b. Australians eat beef.

c. Australians eat crocodiles.

(36a) is a sentence in the original structure and we compare the naturalness of the sentences (36b) and (36c). We would expect (36b) to be more natural than (36c).

Next, we want to look deeper into the actual alternatives that the models generates, because we are ultimately interested in having a model that generates plausible alternatives. We extract a list of most plausible alternatives and filter out the noise. The way we extract alternatives from the model is further explained in Section 3.6. We are interested in how natural these alternatives are. Therefore we investigate how many of the human generated alternatives are in this filtered of most plausible alternatives that is generated by the model. This gives us an indication of whether the model is able to generate alternatives that are, at least partially, in accordance to our human intuitions. We can look at the position the human generated alternatives in this list, which we will refer to as the rank. We would expect a higher rank to make a better alternative. On the other hand, if the alternative follows *even*, it should be a more implausible or unexpected alternative.

The problem, however, is that this rank in itself does not tell us much, since the range of possible alternatives is very large and probably varies substantially per sentence. In order to be able to interpret the ranks, we will need to compare them to the rank of the alternative in a similar condition where the alternative is not unexpected.

Since the alternatives are ordered on a scale, we can investigate this ordering and measure if the models are able to capture this property, induced by *even*. If the models are able to do this, it would be evidence that the models are able to capture at least some of the restrictions put on the alternatives by focus particles, similar to humans. This, in turn, would be further evidence that the models are able to capture human intuitions on alternative sets.

A very easy and straightforward comparison is to remove the word *even* from the sentence and again generate a list of alternatives. We can then compare the ranks of the condition where *even* is included in the sentence to the condition where *even* is excluded from the same sentence. The two conditions are shown in (37).

(37) a. Condition 1: *Subject Verb Object* and even *Alternative*.

b. Condition 2: *Subject Verb Object* and *Alternative*.

Because the effect of *even* is to restrict the set of alternatives to only more implausible alternatives than the object, we would expect the most plausible alternatives of the sentences without *even* to be more varied and thus that the rank of the human alternatives is lower.

In a second experiment, we compare the alternatives of the original sentence with the sentence where we replaced the object with human-generated alternative. The conditions are shown in (38). Note that this experiment is very similar to the experiment where we compare the naturalness of the same conditions, but it differs in that the list of alternatives has been filtered here. Also, we restrict our analysis here to one-word alternatives for both the object and alternative, because we can only easily generate one-word alternatives. And finally, we only isolate the influence the most plausible word that is generated instead of the naturalness of the whole sentence.

(38) a. Condition 1: *Subject Verb Object* and even *Alternative*.

b. Condition 2: *Subject Verb Alternative* and even *Object*.

3.2 Data Collection

3.2.1 Predicting Focus Particles

The sentences are collected through the Sketch Engine platform[19] from the following sources: Wikipedia, ukWaC (Ferraresi et al., 2008), the TenTen English Web Corpus from 2015 (Jakubíček et al., 2013) and the JSI English Timestamped Corpus (Trampuš and Novak, 2012). The sentences are selected manually and are selected, based on their interpretability and occurrences of frequent words. Since we provide these sentences to humans in the form of a judgment task, the sentences must be easily understandable and understandable out of context.

[19]Sketch Engine, `https://www.sketchengine.eu/`

All sentences are of the form Subject-Verb-Object and contain a focus particle either in front of the subject, verb or object. We exclude recursive sentences and each NP consists of at most a determiner, an adjective and a compound noun containing at most two, and at least one noun. Furthermore, we limit sentences to contain only a single verb or a regular verb with an auxiliary or modal verb.

The reasoning behind keeping the complexity of the sentences low, is to make sure the association of the focus particles with the element is more predictable. That is, the focus particles associates only with the element that it precedes. Because focus is not marked in written text, there is no way to resolve potential ambiguities that could arise other than using existing knowledge. Note that there are still potential ambiguities that could arise when the focus particle can associate with multiple elements. Recall that focus particles can associate with every element that it c-commands. For example, when a focus particle is placed in front of an NP that contains an adjective, it could associate with both the noun as well as the adjective. We have also seen that ambiguity arises when the focus particle is placed before the verb. In most cases, these ambiguities can be easily resolved by looking at the surrounding context.

We isolate the sentences on purpose and do not include context around the sentences. We are interested in the first place in the semantic behavior of the focus particles and the extent to which alternatives can be semantically determined. All sentences in the corpus are grammatical when the focus particle is removed and all sentences are grammatical, but certainly not always equally acceptable, when the focus particle is replaced by its counterpart.

In our simple sentences, there are three possible positions to place the focus particle: before the subject (at the beginning of the sentence), before the verb and before the object. For each position and focus particle, we collect 100 sentences, which totals to 6 configurations and thus 600 sentences.

We exclude sentences that contain infrequent words. Because we also feed the sentences to an LM, we want to give the model a fair chance at evaluating the whole sentence. If it contains words it has not seen or words with a low frequency, the sentence might become hard to interpret. Our sentences are short and the infrequent words will most probably be nouns that are crucial to understand the sentence in order to assign the right focus particle. All words in the sentences are therefore searched for in the dictionary of the LMs. This dictionary is created in the training process of the models and discards all words with a frequency of less than 20. If the word is not in the dictionary, we discard the whole sentence. This means that we only keep sentences where the LMs have seen each word in training at least 20 times and have thus constructed a reliable representation of those words.

3.2.2 Generating Alternatives

The sentences for the second part are contructed by hand. We use the simple template shown in (39) as a basis for the human participants, and also for the models to generate alternatives.

(39) *Subject Verb Object* and even ___.

The subject, verb and object vary and could each be composed of multiple words. We only use one template where the alternatives that are generated, are nouns and are in object position.

The sentences are artificially constructed, with the hope that this way we can measure the biggest effects. Besides, we could not find many sentences of this structure in corpora. Some examples of such sentences that have been constructed, are given in (40).

(40) a. Expert servicemen maintain sailing boats and even ___.

b. Each morning starts with coffee and even ___.

c. The restaurant serves oysters and even ___.

In total, we constructed 57 sentences. Some of these sentences share the same object. The original intention was, to also investigate the effects of the different subject and verb on the set of alternatives that is generated, but because of time contraints, this will not be investigated in this thesis.

3.3 Human Judgment Tasks

3.3.1 Predicting Focus Particles

For the annotations, we use the crowdsourcing platform *Figure Eight*.[20] This is a platform where users can complete small tasks and get a small compensation for it. It is an easy and quick way to collect judgments from many participants, in particular native English speakers. In our task, we restrict the participants to those from native English-speaking countries. Concretely, this is realized by restricting the IP addresses of participants to those from the following countries: Australia, Ireland, New Zealand, United Kingdom and United States. Overall, each sentence is annotated by a total of five annotators to make the quality of our annotations reliable. Our corpus consists of 600 sentences. This totals to 3000 judgments that we collect.

The participants have to pass a quiz, consisting of 10 test questions before they can progress to the actual annotation of the sentences. In order to pass this quiz, participants have to get at least 7 of the 10 questions of correct. Furthermore, they are regularly tested on hidden test questions to keep ensuring the quality of the participant's annotations, since we exclude underperforming participants. The test questions are roughly balanced between *even* and *only* to avoid potentially biasing the participants toward choosing one focus particle more often than the other.

To further ensure the quality and also to ensure variety, we limit the maximum number of judgments of each participant to 70. This includes both the test questions

[20] Figure Eight, `https://www.figure-eight.com/`. Until recently, Figure Eight was known as CrowdFlower.

and quiz questions and means that participants can at most judge 54 sentences. In addition, we restrict the participants to only those that had built a reliable reputation on the platform through successfully completing other tasks.

3.3.2 Generating Alternatives

The experimental setup for this task is roughly the same as for the task of predicting the correct focus particle. Again, we use Figure Eight for the annotations. We keep most settings similar and, again, restrict the IP addresses to only those from native English-speaking countries and restrict the participants to the most reliable group annotators.

The corpus for this task consists of 57 sentences and we collect 8 annotations per sentence, because we want to capture variety in the alternatives that participants generate. This totals to 456 judgments. Each participant can maximally complete 20 sentences to ensure the quality and variety of the answers. This is also done because we cannot automatically test the quality of the annotations as there is no restricted set of correct answers.

After the collection of the data we clean up the results. We manually go through the sentences and remove nonsensical answers and correct others for spelling mistakes. In addition, we remove those alternatives that contain words that are not part of the vocabulary of the LMs (e.g., these words have been seen less than 20 times in the training phase of the models). It would not be fair to give the LMs these alternatives, since the model is not able to build a representation for them. We also remove all answers that are not nominal alternatives (such as verbal alternatives). In this task, we restrict our investigation to only nominal alternatives. Finally, we remove duplicate answers where two people gave the same answer.

3.4 Computational Models

To computationally model the human intuitions on alternatives, we use two recurrent neural networks (RNNs). More specifically, we use a long short-term memory (LSTM) and the bidirectional version of an LSTM, which we will refer to as BiLSTM. Recall that recurrent neural networks process a sentence from left to right and thus match the human sentence processing well. This means they have at least some cognitive plausibility. They have also been among the most successful models in language modeling tasks (Jozefowicz et al., 2016; Mikolov et al., 2010). LSTMs have also been shown to capture longer-distance dependencies, because they are able to selectively forget and remember information (Bowman et al., 2015; Dai et al., 2019; Kuncoro et al., 2018; Trinh et al., 2018).

We use a slightly modified version of the LSTM model that was used for judging the grammaticality and acceptability of a sentence in (Chowdhury and Zamparelli, 2018). We also use a slightly modified bidirectional version of this first LSTM model as our second model. We use the models as provided to us and do not alter any settings

ourselves. Note that the BiLSTM model does not process sentences incrementally, because it also processes the sentence from right to left. BiLSTMs are capable of learning more complex patterns as at every time step they incorporate information from both sides of the current word that is processed, instead of only information from previous words in the sentence.

For both models, each word is encoded using a one-hot vector encoding. This means that the length of the vector is as large as the vocabulary size and that each vector consists of zeros and only a single one in the position that is unique for the word it is encoding. Next, these one-hot vectors are converted to dense embedding vectors of size 256 each. These dense embedding layers serve as input for the hidden layers. Both models use 2 hidden layers with each 500 hidden units for the LSTM model and 250 hidden units for the BiLSTM model.

In both models, regularization is applied in the form of dropout. This means that some hidden units are randomly not updated in the training phase, in order to avoid overfitting. Furthermore, a batch size of 80 is used in both models and 25 epochs are used in the LSTM model, while only 3 are used in the BiLSTM model. A batch size of 80 means that in each iteration, 80 sentences were trained on at the same time. 1 epoch means that all sentences have been used in training once. Both models were implemented in PyTorch and optimized using stochastic gradient descent.

The models are not optimized or fine-tuned for the best performance. Since the original purpose of the models was to only show they are able to learn syntactic patterns, they were not optimized. For our task, we can make a similar argument. We are exploring the potential of neural LMs to capture human intuitions on alternatives. We are thus not searching for the best performing model, but only exploring the ability of LMs to capture human intuitions on alternatives

The models are trained on a corpus consisting of Wikipedia texts, (English) novels from the Project Gutenberg corpus[21] and general British web pages. The total corpus consists of approximately 31 million sentences or 0.7 billion words. The vocabulary size is approximately 100,000 words. From this vocabulary, the words with a frequency of less than 20 are replaced with an *unknown* token and possibly a suffix giving some clues on the type of word. The sentences are further preprocessed by adding tags indicating the start and end of the sentence, and all numbers are replaced by a single token. Finally, all upper case letters are converted to lower case.

3.5 Evaluation Measures

The evaluation measures are used to measure how predictable a sentence is in a given model. This gives us some indication of the naturalness, acceptability or likelihood of a sentence, as perceived by an LM. Thus, the evaluation measures assign a score to a sentence, stating how acceptable, natural or plausible the sentence is. These evaluation measures are furthermore used in deciding the most plausible focus particle. We will first discuss SLOR and then explain perplexity.

[21] Project Gutenberg, `https://www.gutenberg.org/`

3.5.1 SLOR

The syntactic log-odds ratio (SLOR) is a metric that has been proposed to measure the grammaticality of a sentence. SLOR is a relatively new measure. The use of SLOR for measuring the acceptability of a sentence was first proposed by Pauls and Klein (2012). Lau et al. (2016) investigated the correlation between human grammaticality judgments and the scores assigned by different evaluation measures, including SLOR, on a variety of LMs and found that SLOR was the best measure to calculate the acceptability of a whole sentence on global level.

It is important to note that in their experiments, Lau et al. (2016) were only interested in the grammaticality of a sentence. This means that it is not entirely clear if this metric can also measure how semantically acceptable a sentence is. In our experiments, we are only interested in testing how semantically acceptable or how natural a sentence is, as all our sentences are grammatical. Because semantically unacceptable sentence will also not be part of the corpora, we can assume the LMs to also learn patterns of semantic acceptability, next to patterns of syntactic acceptability. Because SLOR measures the likelihood of a sentence of occurring in the data, we expect that it should be capable of capturing semantic acceptability, just like it is able to capture syntactic acceptability.

SLOR consists of three components: the log probability of the sentence as given by the language mode, the unigram probability of the sentence and finally the sentence length. Essentially, the probability that the model assigns to the sentence is corrected for both frequency biases and a bias toward preferring shorter sentences. The unigram probability of the sentence is used to filter out frequency biases of individual words. If this was not done, SLOR would tend to assign higher scores to sequences with more frequent words. Note however, that filtering out frequency biases only occurs at a global level. There could still be a frequency bias when looking at a specific position in a sentence. The sentence length is used to normalize the sentence. The normalization is done to prevent longer sentences from being perceived as less natural.

The log probability of the sentence is in our case defined as the inverse of the *cross entropy*.[22] We will continue to use the term cross entropy, because in the next paragraph we will see that perplexity also makes use of it. Besides, there is the methodological argument for using this term: cross entropy is easier to use in explaining how the LM calculates this probability. Cross entropy is a measure that measures how well one probability distribution can explain another. In terms of neural networks, it is measured by calculating the distance between the one-hot encodings for each word in the sentence and the output distribution of each word. This can be approximated with Equation (3.1) that calculates the cross entropy for a sentence ϕ.

$$H(P, \phi) = -\sum_{i=1}^{|\phi|} \ln P(\phi_i|\phi) \tag{3.1}$$

In Equation (3.1), P is the probability distribution over a sequence of words and is

[22]Note that this is not always the case and only holds because we are dealing with sequential data.

computed by the LM. ϕ is the sequence of words for which we measure the probability and ϕ_i is the ith word in this sequence. Note that the probability of a word in the sentence depends on the sentence. In the LSTM model, concretely this means that the model depends on all words on the left of the ith word. In the BiLSTM the ith word depends on all other words. Note further that the base that is used, is the natural number e.

The unigram count P_u is the product of all unigram probabilities as is shown in Equation (3.2).

$$P_u(\phi) = \prod_{i=1}^{|\phi|} P(\phi_i) \tag{3.2}$$

These unigram probabilities do not depend on the sentence and are measured as the frequency of occurrence on the whole corpus the model was trained on.

The full equation is given in (3.3). As mentioned before, the log probability is the inverse of the cross entropy. The unigram probability is subtracted from the log probability and finally the score is normalized by the sentence length.

$$SLOR(P, P_u, \phi) = \frac{-H(P, \phi) - \ln(P_u(\phi))}{|\phi|} \tag{3.3}$$

3.5.2 Perplexity

Perplexity is a metric that measures how well a sequence of words or sentence is predicted by an LM. If the LM is more perplexed (indicated by a higher score), it means that the sentence was less predictable and thus less natural. Like with SLOR, it is measured using cross entropy as was defined in Equation (3.1). The perplexity is then defined as in Equation (3.4).

$$PPL(P, \phi) = e^{\frac{H(P,\phi)}{|\phi|}} \tag{3.4}$$

Here, the cross entropy is normalized by the sentence length $|\phi|$ and this number is exponentiated. Like with SLOR, the sentence is normalized by the sentence length to prevent biases toward shorter sentences. Unlike SLOR, perplexity does not correct for frequency differences of individual words.

3.6 Extracting Alternatives

Because the LMs compute the likelihood of a word, given a sequence of words, we can also compute the likelihood of every other word in the vocabulary in a specific position. By ranking all words in the vocabulary on their likelihood, we can extract the most plausible words in that specific position. Note however, that this only lets us generate one-word alternatives. If we would want multiple word alternatives, we need to define a decision mechanism that decides when to generate more words. One relatively easy way is to use specific part-of-speech (POS) tags as cues. Such POS tags could be

adjectives or determiners, which are normally followed by nouns. In the interest of time, we restrict our analysis to only single-word alternatives

In our experiments, we always generate the 1000 most plausible words. This number is chosen, because there is normally a lot of noise in this list, which we need to filter out. After the generation of the alternatives, we filter out all tokens that replaced infrequent words or numbers in the training phase. Furthermore, we filter out words that are not made up of letters, such as punctuation symbols.

Since we only restrict alternatives to single words, we want these words to be nouns. We thus need a way to assign a POS tag to the word. There are two possible ways to do this. The most straightforward way would be to give the whole sentence, including the alternative generated, to a POS tagger that assigns a POS tag to each of the words, based on the sentential context. Another way is to look at each word out of context and assign all possible tags to the word.

We implement the second method and do not experiment with the first. The main reasons for choosing this approach are, that it did not require an optimized POS tagger, which still could potentially assign a wrong POS tag to the word. We first lemmatize each word. Next, we use the lemmatized word and retrieve all possible tags from WordNet for it. WordNet is a large and very complete lexical database for English (Miller, 1995). Each lexical entry is assigned all possible tags. The tags are not very fine-grained and only consists of the following classes: nouns, verbs, adjectives and adverbs. Since we are only interested in nouns in general, we do not need more fine-grained information and thus this tagset is sufficient for our purpose. Finally, we select all words that were assigned a noun tag for our list of alternatives.

Chapter 4

Results

4.1 Human Judgments

4.1.1 Predicting Focus Particles

In total, we had 64 persons participating in the human judgment task of predicting focus particles. Of those, one failed the quiz and was excluded from the experiment. Two participants only completed the quiz and did not complete any judgments. This means that we had a total of 61 participants that judged 3000 sentences and that each participant on average completed 49.18 judgments.

We calculated the inter-annotator agreement for all sentences using Fleiss' kappa. The overall agreement was 0.68. This means that there was substantial agreement. This, in turn, gives us an indication that, at least in our corpus, containing only short and simple sentences, humans have fairly consistent intuitions on which focus particle best fits in a certain context. We also calculated the inter-annotator agreements for each syntactic position separately. The inter-annotator agreement scores have been summarized in Table 4.1.

From the inter-annotator agreement scores, we can observe, that the agreement between judgments where the focus particle was in object position, was a substantially higher than average. One potential explanation for this might be, that sentences are processed incrementally (Kamide et al., 2003; Tanenhaus and Brown-Schmidt, 2008; Tanenhaus et al., 1995; Traxler et al., 1997). The the set of alternatives of an element in object position could be more constrained and therefore easier to judge than the set of alternatives of elements in subject or verb position. That is, because the humans have more information about the sentence available to make the judgment in object

	S	**V**	**O**	**Overall**
Agreement	0.67	0.56	0.81	0.68

Table 4.1: The Fleiss' kappa inter-annotator agreement scores, split per syntactic position ((S)ubject, (V)erb and (O)bject).

	S	**V**	**O**	**Total**
even snts judged as *even*	92	94	96	282
only snts judged as *even*	8	19	2	29
Total *even* snts	**100**	**113**	**98**	**311**
only snts judged as *only*	92	81	98	271
even snts judged as *only*	8	6	4	18
Total *only* snts	**100**	**87**	**102**	**289**

Table 4.2: The distribution of how the sentences were judged. The distribution has been split per position ((S)ubject, (V)erb and (O)bject position) and per category.

position.

From the agreement scores, we can also note that there was substantially less agreement for sentences in verb position. These sentences are thus more ambiguous in the alternatives they induce and therefore harder to judge. The most plausible cause of this effect is that focus particles in verb position could associate with both the verb and the object, making the set of alternatives, and with that the sentence ambiguous, resulting in more variation and hence, more disagreement in the judgments. Another potential explanation is that alternatives of verbs might be harder to compute for humans than alternatives of nouns. There seem to be different processing mechanisms for nouns and verbs in the brain (Breedin et al., 1998; Damasio and Tranel, 1993; Damasio et al., 1996; Mätzig et al., 2009). Nouns are easier to process in general than verbs (Abel et al., 2015; Masterson et al., 2008) and are clearer organized in classes (Huttenlocher and Lui, 1979; Vigliocco et al., 2011). Therefore, it might be easier to compute alternatives for nouns than for verbs.

The judgments were fairly even distributed, with 51.7% of all 3000 judgments being *even* and 48.3% of the judgments being *only*. This means that the participants did not have an overall bias toward the use of one of the focus particles. The distribution of the judgments by the participants is summarized in Table 4.2. Recall that originally there were 100 sentences per focus particle and syntactic position. The sentences have more or less been equally distributed between *even* and *only*. We can observe that for both *even* and *only*, most sentence judgments corresponded with the original focus particle. The fact that there were so few sentences in object position that changed focus particles is consistent with our hypothesis that sentences in object position might be less ambiguous, and thus easier to judge.

One notable exception is that quite a few sentences that originally had *only*, were now judged as *even*, leading to a slight imbalance. As observed previously, there was more disagreement for focus particles in verb position and they seem harder to judge. The participants also might have a slight bias toward choosing *even* in verb position, although a quick search in the corpus shows that *only* is about three times more likely than *even* in verb position. Combining these two observations, this effect could be explained by the bias becoming more pronounced as a consequence of the difficulty of judging sentences with the focus particle in verb position.

	S	**V**	**O**	**Total**
even snts judged as *even*	60	45	81	186
only snts judged as *even*	0	1	2	3
Total *even* snts	**60**	**46**	**83**	**189**
only snts judged as *only*	71	63	75	209
even snts judged as *only*	1	2	0	3
Total *only* snts	**72**	**65**	**75**	**212**

Table 4.3: The distribution of how the unambiguous sentences were judged. The distribution has been split per position ((S)ubject, (V)erb and (O)bject position) and per category.

(41) gives one example sentence for each focus particle in each position. The sentences are shown with the particle that was chosen the most by the participants.

(41) a. **Even** the car industry uses bicycles.

b. The company **even** adopted a new logo.

c. The speech surprised **even** the Russian delegation.

d. **Only** subscribers have full access.

e. Substance abuse **only** perpetuates a continuous cycle.

f. These methods yield **only** probabilities.

(41b), (41c), (41e) and (41f) did not have full agreement on the focus particle. As can be noticed, those sentences are ambiguous in that *even* could be replaced by *only* or vice versa. Finally, (41a) and (41d) are not very acceptable, if their counterpart would be used instead.

This ambiguity could also become a problem for the LMs, as both focus particles might be equally plausible in some sentences. For this reason, we excluded all sentences where there was not full agreement. In Table 4.3, the distribution of the unambiguous sentences is outlined. In unambiguous sentences, there was full agreement by five annotators on the judgment of the focus particle.

In total, there were 401 unambiguous sentences, which were roughly evenly distributed between sentences judged with *even* and sentences judged with *only*, with only slightly more sentences unambiguously judged as *only*. In the next section, we will use these judgments to investigate how well the computational language models are able to capture these similar intuitions.

4.1.2 Generating Alternatives

For the experiment where we asked participants to come up with alternatives, we had 26 participants. With 456 judgments in total, this means that each participant, on

average, completed 17.54 judgments. After cleaning up the answers manually and removing duplicates, nonsensical responses and answers where the models did not have a representation of, because the words were seen too infrequent in the training phase, we were left with 315 answers. In the experiments where we have the model generate alternatives, we limited these answers to only one-word alternatives. There were 209 one-word alternatives, distributed over 54 different templates (e.g., a sentence without the alternative).

We analyzed the generated alternatives and found that they were often not as extreme or unexpected as we had anticipated. In many cases, we were able to switch the object and alternative, while the sentence would still be natural, and in some cases even more natural than the original sentence. These factors complicated our analysis, because part of it is based on measuring these effects. Our setup was meant to generate a more unexpected alternative, so that we could compare the alternative to the object, using our models and draw conclusions, based on this.

In (42), we show some examples of alternatives generated by the participants. All examples were taken from the list of sentences after filtering. If we switched the object and the alternatives, (42b), (42c) and (42g) would be perfectly fine in our intuitions and (42c) sounds even more natural when the object and alternative are switched, because we expect to find crabs on the menu in a restaurant more often than oysters. Switching the object and alternative in (42a) is not unthinkable either, in the right context, although we could argue that the sentence would become less plausible and less natural. Perhaps, we are talking about a very exclusive harbor, such as the one from Monaco where we would only expect yachts and not sailing boats. In (42d), (42e) and (42f) on the other hand, it seems hard to switch the object and the alternative.

(42) a. The harbor houses sailing boats and even ___. – Alternative: yachts

b. Expert servicemen repair watches and even ___. – Alternative: televisions

c. The restaurant serves oysters and even ___. – Alternative: crabs

d. The art lover appreciates paintings and even ___. – Alternative: sculptures made from recycled materials

e. The beach attracts tourists and even ___. – Alternative: the occasional sun bathing seal

f. Rich people buy villas and even ___. – Alternative: castles

g. The old lady has lived in Italy and even ___. – Alternative: Greece

The alternatives in (42c) and (42g) do not seem very extreme when comparing them with the objects of the corresponding sentences. In fact, *oysters* would seem a more extreme alternative than *crabs* to serve in a restaurant in (42c). The reason why *even* still seems to work in those sentences, is perhaps, that, again, it influences the expectations of the subject and through that it might change the scale of *even*. For

example, we might have an expectation that the restaurant serves either oysters or crabs, if it is specialized in either one of them. The other alternative in that case is always more unexpected.

It seems that switching the object and the alternative not only changes or reverses the scale, but often also changes our expectations about the subject and/or verb. Because our sentences do not have any context, scale changes might be more apparent. This had severe consequences for our further analysis where we assumed the generated alternative to be more implausible or unexpected than the object. Since this is an exploratory study, we did not have expectations on how the results would be. For this reason, we continued our analysis as planned, but we took into account that our data might not be good enough to observe differences between conditions.

4.2 Predicting Focus Particles

This section is divided into four parts and is aimed at understanding how well the LMs are able to capture human intuitions on focus particles. We start by analyzing the naturalness of sentences with the correct and incorrect focus particle, as judged by humans. After that, we investigate the isolated contribution of the focus particles to the overall naturalness of the sentences. We then investigate the performance of the models at predicting the correct focus particles, using the human judgments as gold standard to compare to. We close this section with a brief discussion.

4.2.1 Naturalness of the Sentences

We will start our analysis by looking at how natural or acceptable the model perceives the sentences. We used the sentences with focus particle judgments as gold standard, and only used the unambiguously judged sentences. Thus, sentences that the participants judged as *even* should in principle be more natural than the same sentence, but with *even* replaced by *only*. Likewise, we would expect sentences that the participants judged as *only* to be more natural than the same sentences, but with *only* replaced by *even*.

We measured the naturalness of the four conditions separately and also measured them per syntactic position. Finally, we averaged the scores from the different syntactic positions. We evaluated the naturalness of the sentences on both the LSTM and BiLSTM models and measured this naturalness using both SLOR and perplexity. We will refer to sentences, judged unambiguously as *even*, as "*even* sentences". Likewise, to sentences, judged unambiguously as *only*, we refer as "*only* sentences".

Table 4.4 summarizes the results of the evaluation on the LSTM model, using SLOR. Table 4.5 summarizes the same results, but measured using perplexity. In these tables, we are only allowed to compare sentences with the correct, as judged by the participants, to the same sentences with the incorrect focus particle. We are not allowed to compare *even* sentences with *only* sentences, because the sentences are different, apart from the focus particle. The same holds for comparing across different

	S	V	O	Avg per FP
even snts with *even*	1.219	1.403	1.495	1.372
even snts with *only*	1.193	1.498	1.330	1.340
even snts without FP	1.234	1.490	1.360	1.361
only snts with *only*	1.484	1.609	1.699	1.597
only snts with *even*	1.465	1.460	1.562	1.496
only snts without FP	1.478	1.527	1.651	1.552

Table 4.4: The overall SLOR values for sentences with the correct and incorrect focus particle, as judged by the participants, split per syntactic position ((S)ubject, (V)erb and (O)bject). We have also included the SLOR values for the same sentences, but without a focus particle (FP). The evaluation was done using the LSTM model.

	S	V	O	Avg per FP
even snts with *even*	192.4	183.4	165.2	180.3
even snts with *only*	183.4	159.0	179.6	174.0
even snts without FP	168.5	157.7	169.2	165.1
only snts with *only*	150.5	156.1	94.3	133.6
only snts with *even*	164.4	291.7	113.4	156.5
only snts without FP	142.2	167.5	91.4	133.7

Table 4.5: The overall perplexity scores for sentences with the correct and incorrect focus particle, as judged by the participants, split per syntactic position ((S)ubject, (V)erb and (O)bject). We have also included the perplexity scores for the same sentences, but without a focus particle (FP). The evaluation was done using the LSTM model. Note that lower perplexity scores mean a better performance.

syntactic positions.

Let us first examine the average SLOR values of all syntactic positions in Table 4.4. We can observe that for sentences with the correct focus particle, as judged by the participants, the SLOR values are higher than for the same sentence, but with the focus particle replaced. This is in accordance with our expectations that sentences that were judged with the correct focus particle are more natural than the same sentences with the incorrect focus particle. This is a first indication that the model is capable of correctly distinguishing between correct and incorrect uses of focus particles and are thus learning something about the alternative sets.

Both in the case of *even* sentences and *only* sentences in object position, there is quite a large difference between the correct and incorrect predictions. The correct prediction is, indeed, in general seen as more natural. This large difference is consistent with the fact humans also found focus particles in object position easier to judge. So, for both the participants and the models, it seems easier to judge the sentence on the most plausible focus particle when it occurs in object position. In the case of the model, one explanation could be the way the LSTM model processes the sentences, which is

from left to right. This means that in object position, the most information about previously seen words, is available to the model to use in the evaluation.

Both in the case of *even* sentences in verb position and *only* sentences in subject position, the sentences with the incorrect focus particle are seen as more plausible. The predictions in subject position could be explained by a bias toward *even*. Because the LSTM model processes the sentence incrementally (from left to right) and the focus particle in subject is the first word in the sentence, there is no information to base the decision on. This means that the model has to rely solely on frequency information on a focus particle, occurring in subject position. A quick search on Sketch Engine reveals that *even* is about three times more plausible in subject position, relative to other syntactic positions, than *only*, relative to other syntactic positions. Biases related to preferences of a focus particle for a specific syntactic position are not corrected for in SLOR, which could explain our findings.

A comparison of the sentences with a focus particle to the same sentences without any focus particle in Table 4.4, reveals that sentences with the correct focus particle are perceived as more natural by the model than the same sentences without any focus particles. We can also observe the opposite effect: sentences with the wrong focus particle seem to be judged by the model as less natural than the same sentences without a focus particle. This seems to suggest that sentences become more natural if the correct focus particle is used, and that not always both focus particles make the sentence equally natural, although it does not have to be the case for every focus particle. This is another piece of evidence, pointing toward the model being able to learn about alternative sets, since the difference between the contribution of *even* and *only* is the way they restrict the alternative set of the element they associate with. The *only* sentences seem also more natural in general than the *even* sentences, if we remove the focus particle. This could explain the fact that in general the values for *only* sentences are higher than for *even* sentences.

The results on perplexity in Table 4.5 look quite similar, but seem more biased toward preferring *only* over *even*. First, note that lower perplexity scores mean that the model is less surprised, which therefore means that the sentence is more natural, according to the models. It could be the case that there is a bias toward *only*, because perplexity does not account for global frequency biases. *Only* is more frequently used in English and thus most likely also more frequently seen by the model in training. This bias seems to result in the tendency of the model to perceive *only* sentences as more natural in general and leads to incorrect expectations on the naturalness of *even* sentences in both subject and verb position.

The scores in object position seem less biased toward *only*, as in those cases the perplexity scores for the sentences with the focus particle, as judged by the participants, are correctly predicted to be lower. This might be explained by the fact that the model has more information on which to base the judgment and does not have to rely only on frequency information of occurring in a specific syntactic position. This might filter out the bias.

Finally, we can observe that the sentences without the focus particles are generally

	S	V	O	Avg per FP
even snts with *even*	3.901	4.185	4.401	4.162
even snts with *only*	3.953	4.167	4.408	4.176
only snts without FP	3.649	3.798	3.976	3.808
only snts with *only*	3.990	4.170	4.190	4.116
only snts with *even*	3.950	4.165	4.167	4.094
only snts without FP	3.725	3.776	3.714	3.738

Table 4.6: The overall SLOR values for sentences with the correct and incorrect focus particle, as judged by the participants, split per syntactic position ((S)ubject, (V)erb and (O)bject). We have also included the SLOR values for the same sentences, but without a focus particle (FP). The evaluation was done using the BiLSTM model.

perceived as more natural by the model when they are evaluated using perplexity, as opposed to the model evaluating the sentences using SLOR. Even though perplexity does compensate for differences in sentence length, it still seems that shorter sentences (e.g., sentences without the focus particle) are generally perceived as more natural.

The results for the BiLSTM are outlined in Table 4.6, for the evaluation using SLOR. The first thing to notice is that the SLOR values are a lot higher in general. This must be the effect of the bidirectionality, as this allows the model to learn more complex patterns from the data. That is because the processing happens not only from left to right through the sentence, but also from right to left.

As we can observe, the *even* sentences with the incorrect focus particle are, contrary to our expectations perceived by the BiLSTM model as more natural, although the difference is small. For *only* sentences, the correct predictions are made, but the differences are not so large as with the LSTM model. From these results, it seems that the BiLSTM model has more difficulty in distinguishing between sentences with the correct and incorrect particle as judged by the participants. Again, there seems to be a small frequency bias toward *only*, since sentences with this focus particle are more natural according to the BiLSTM model, regardless of whether *only* was correctly or incorrectly used.

The differences in SLOR values between sentences with the correct focus particle and sentences with the incorrect focus particle are smaller on the BiLSTM than on the LSTM model, even though the overall values are a lot higher. These smaller differences between different conditions might be caused by the fact that the model is optimized for predictability. The model is more complex, since it also processes the sentence from right to left and can therefore make more precise predictions. Therefore, the model could have lost the ability to distinguish between such subtle differences that are due to the focus particle, because the sentences with any focus particle might be perceived as equally natural by the model. This is called a *ceiling effect*.

Furthermore, contrary to the LSTM model, the sentences without focus particles are perceived as less natural than sentences with either the correct or incorrect focus particle. We do not know what human intuitions would be on this, but it is interesting

to note this difference between the LSTM and BiLSTM model using the SLOR evaluation. We expect that it is very sentence-dependent whether a sentence would be more or less natural when a focus particle is used, as compared to the sentence without a focus particle. The LSTM model using perplexity also seems to also prefer sentences without a focus particle over sentences with a focus particle.

The results for the BiLSTM, using perplexity, are not summarized in a table here. For all comparisons between the sentences with the correct and incorrect predictions, the sentences with *only* were perceived by the model as more natural. This again clearly indicates a bias toward *only* as the perplexity metric does not compensate for frequency biases.

Overall, the scores on perplexity for the LSTM model seem to suggest that when it is evaluated using perplexity, it is less able to capture the human intuitions on focus particles than when the model is evaluated using SLOR. We also correlated the scores from the evaluation using SLOR with the perplexity scores. There is a correlation between both metrics that both are supposed to measure the naturalness of a sentence. The Pearson's correlation coefficient is $r = -0.5237$ for the LSTM model and $r = -0.6038$ for the BiLSTM model. Because of this correlation we will not continue to always report the scores for both evaluation metrics. Instead, we will focus on reporting the SLOR values, because SLOR seems better at measuring the naturalness and less influenced by frequency biases, as we will also see later.

We will now summarize the most important findings from this section. First of all, the SLOR evaluation metric seems to correspond better with the human intuitions on naturalness of sentences with focus particles than perplexity. This is most probably the case, because perplexity seems to be too much influenced by frequency biases. Secondly, the LSTM model, in contrast to the BiLSTM model, seems better at capturing the differences between *even* and *only*, since the differences in naturalness as measured by the BiLSTM model seem smaller when comparing to the differences in naturalness using the LSTM model. This is perhaps surprising, given the fact that the BiLSTM should capable of learning more complex patterns from the data. The overall naturalness of the sentences is also a lot higher on the BiLSTM model.

4.2.2 Contribution of the Focus Particles to the Sentences

To better understand whether the contribution of a focus particle differs in correct and incorrect contexts, we tried to isolate the exact contribution of the focus particle. We only computed this for the models that use the SLOR evaluation metric as the models that use perplexity do not seem to capture the human intuitions as well as SLOR seems to do. We isolated the exact contribution of the focus particle for a specific sentence by subtracting the SLOR value for the sentence without a focus particle from the same sentence with either the correct or incorrect focus particle. The averages on the sentences for the LSTM model using SLOR are shown in Table 4.7. These results have, like before, been split per syntactic position

In Table 4.7, we compare the contribution of the correct focus particle to the contribution of the same focus particle in the incorrect sentence. The results are averaged

	S	V	O	Avg per FP
Contribution of correct *even*	-0.014	-0.086	0.135	0.011
Contribution of incorrect *even*	-0.013	-0.068	-0.089	-0.051
Contribution of correct *only*	0.006	0.082	0.048	0.045
Contribution of incorrect *only*	-0.041	0.009	-0.030	-0.021

Table 4.7: The isolated contribution of the correct and incorrect focus particle to the sentence as evaluated by the LSTM using SLOR.

over all sentences for each condition. Because we subtract the contribution of the sentence itself, we are directly comparing the effect of the focus particle on the sentence between its correct and incorrect use. We would expect the correct focus particle to have a more positive impact on the SLOR values than the same focus particle used in an incorrect sentence. This indeed seems to be the case for most conditions. For both *even* and *only*, there is a difference, where the correct prediction has a more positive contribution. Moreover, the difference in contribution between *even* in correct and incorrect contexts is significant ($p < 0.001$). Similarly, the difference in contribution between the *only* in correct and incorrect contexts is also significant ($p < 0.001$).

The differences for incorrect focus particles tend to be negative, which, as explained before, means that the sentences are perceived as less natural, as compared to the sentences without the focus particle. Furthermore, we can observe that *only* generally seems to have a larger positive contribution to the sentence than *even*. This might be caused by the frequency bias toward *only* that does not seem to be completely filtered out.

Furhtermore, the contribution of *even* in correct contexts in subject and verb position is negative and in verb position even more negative than the incorrect use of *even*. We do not have an explanation for this observation. In all other conditions than *even* in subject and verb position, the contribution of the focus particle in the correct context is larger. This is clear evidence that the model is able to distinguish between alternative sets, induced by *even* and *only*.

The same evaluation on the BiLSTM is shown in Table 4.8. First, observe that the contribution of the focus particle is always positive, because, as we have seen previously, the sentences without a focus particle are perceived as less natural by the BiLSTM model. We can observe that the overall differences in contribution between the correct and incorrect focus particle are very small for both *even* and *only*. The differences between the correct and incorrect uses of both focus particles was not significant in any of the comparisons.

To summarize this section. From a closer inspection of the contribution of *even* and *only* on the naturalness, the LSTM model, in contrast to the BiLSTM model, is able to learn correctly the difference between the two focus particles. This further strenghtens the hypothesis that the LSTM model is able to learn about alternative sets, following human intuitions to some extent.

	S	V	O	Avg per FP
Contribution of correct *even*	0.252	0.387	0.425	0.354
Contribution of incorrect *even*	0.225	0.390	0.454	0.356
Contribution of correct *only*	0.265	0.394	0.476	0.378
Contribution of incorrect *only*	0.304	0.369	0.432	0.369

Table 4.8: The isolated contribution of the correct and incorrect focus particle to the sentence as evaluated by the BiLSTM using SLOR.

	S	V	O	Avg per FP
even sentences	69.44	18.46	85.33	57.74
only sentences	53.33	89.13	91.16	77.87
Avg per position	61.39	53.80	88.25	67.81

Table 4.9: The accuracy on the model's judgments of the most plausible focus particle in each sentence, split per syntactic position and per focus particle. Per focus particle, in this case, means the correct focus particle as judged by our participants. The evaluation was done on the LSTM model using SLOR.

4.2.3 Ability of Predicting the Most Plausible Focus Particle

Now we move on to the actual individual predictions of the models on the judgment task. We report the accuracy of how often the model predicts the same focus particle as the participants did. The decision the model had to make, is always between *even* and *only*. This decision is made by comparing the overall naturalness of the sentences with *even* and with *only*. The results for the accuracy on the LSTM model using SLOR for the evaluation is shown in Table 4.9. With an accuracy of 67.81%, the overall accuracy seems to indicate the model has been able to capture some human intuitions. However, this seems mostly due to the good performance in object position. Also, as we predicted from the analysis of the naturalness of the sentences, there is a bias toward *only*.

The performance of the LSTM model on *only* sentences is much better than on *even* sentences. This is especially apparent for the verb position. One exception is subject position, where the performance on *even* sentences was much better than on *only sentences*. This is probably due to a positional bias toward *even* at the start of a sentence.

The accuracy is the highest when the focus particle is in object position and is much higher than chance (50%). Earlier, we also observed that object position was the easiest position for participants to judge, possibly because the set of alternatives is most constrained in this position as most information to make the judgment is available. This could also hold for the LSTM model. However, using the same reasoning we would expect the accuracy on the verb position to be higher than the accuracy on subject position, but we do not observe this. Both the accuracy on subject position as well as verb position is not much higher than chance.

Table 4.10 summarizes the performance of the LSTM, similar to Table 4.9, but

	S	**V**	**O**	**Avg per FP**
even sentences	27.78	6.15	76.00	36.04
only sentences	75.00	95.65	100.00	90.22
Avg per position	51.39	50.90	88.00	63.43

Table 4.10: The accuracy on the model's judgments of the most plausible focus particle in each sentence, split per syntactic position and per focus particle. Per focus particle, in this case, means the correct focus particle as judged by our participants. The evaluation was done on the LSTM model using perplexity.

	S	**V**	**O**	**Avg per FP**
even sentences	18.06	75.38	56.00	49.81
only sentences	83.33	36.96	56.63	58.97
Avg per position	50.70	56.17	56.32	54.39

Table 4.11: The accuracy on the model's judgments of the most plausible focus particle in each sentence, split per syntactic position and per focus particle. Per focus particle, in this case, means the correct focus particle as judged by our participants. The evaluation was done on the BiLSTM model using SLOR.

using perplexity. Generally, we can observe the same tendencies as on the evaluation that used SLOR, but the preference for *only* is more pronounced. The difference in accuracy is most pronounced in verb position, where the model's preference for *only* is very apparent. But also for subject and object position, there is a very large difference between sentences with *even* and *only*. Overall the accuracy is slightly lower than when the model is evaluated using SLOR. Still, also the LSTM model evaluated using perplexity seems to be able to capture some human intuitions on focus particles.

Table 4.11 summarizes the results for the BiLSTM model using SLOR. Overall, the BiLSTM model does not seem able to capture the human intuitions on the difference between *even* and *only*. The BiLSTM model performs only slightly above chance level, while still having some positional biases. This is in contrast to the LSTM model, which does seem able to be able to sometimes correctly distinguish between *even* and *only*, in at least object position. This could be due to the smaller differences in scores assigned to the sentences between *even* and *only*. It is interesting to note that there does not seem to be an overall frequency bias toward *only* in the BiLSTM model, in contrast to the LSTM model.

The results of the evaluation of the BiLSTM model using perplexity are not shown here. But analogous to the comparison between the evaluation using SLOR and perplexity on the LSTM model, the results for the BiLSTM model using perplexity also show a very strong bias toward preferring *only*.

To summarize the most important findings in this section. The LSTM model seems able to capture some human intuitions on focus particles, in at least object position, in contrast to the BiLSTM model which does not seem to be able to at all. Furthermore,

the models seem very sensitive to local and global biases. The results also seem heavily affected by a frequency bias toward preferring *only*, with the BiLSTM using SLOR being an exception. This bias is even more pronounced when the models are evaluated using perplexity.

4.2.4 Discussion

We investigated whether simple LMs are able to capture similar intuitions on focus particles as humans do. We used three analyses, two different LSTM models and two different evaluation metrics. Overall, only the LSTM model, using SLOR for the evaluation, shows consistency in the ability of capturing some human intuitions on focus particles on all three analyses, although the performance is not very high.

When looking the performance on the different syntactic positions, the LSTM model performs best when the focus particle is in object position. This seems to be the easiest position to evaluate for both the human participants and also the LSTM model. For both humans and the model, we argued that this is the position that is the easiest to judge, because both humans and the LSTM model process a sentence incrementally and have most information available to use in the judgment of the most plausible focus particle. Besides, it is easier to compute a set of alternatives for nouns than for verbs (Abel et al., 2015; Masterson et al., 2008).

The BiLSTM model, in contrast to the LSTM model, does not seem to be able to capture any differences between *even* and *only*. On all three analyses, we did not observe a clear indication that the BiLSTM model was able to capture human intuitions on focus particles. We suspect this is due to a *ceiling effect* where the model is too optimized on predictability to measure subtle differences, caused by the focus particle.

The performance on the task of predicting the most plausible focus particle is not very high. There could be multiple reasons for this that are not necessarily mutually exclusive. First, the models have not been optimized. We have not experimented with different parameter settings. A second explanation could be that SLOR is still not a good evaluation metric for measuring semantic acceptability. We have seen it is very sensitive to biases of focus particles toward a specific syntactic position. At least SLOR is better at capturing the human intuitions on focus particles than perplexity. We might still need better evaluation metrics than SLOR if the aim is to find the best performing models on capturing semantic information, and more specifically, information about alternative sets. Furthermore, although the inter-annotator agreement was quite high, it still seems to be a difficult task. This could be explained by the implicit nature of the alternative sets. The models have to rely on explicit information to compute the implicit set of alternatives. Finally, the observed differences between sentences with *even* and *only* might completely be based on the association with words in the context. Some words might be more likely to co-occur with one focus particle than the other. We can therefore argue that the models do not learn anything about focus particles and alternative sets, but simply learn to associate some words better with one focus particle.

Nevertheless, this co-occurrence information is still relevant information for the

model in learning about alternative sets and we argue that this information is part of the definition of learning about alternative sets. If an element always co-occurs *only* in a specific context, we might infer from this that, this element does not occur in combination with other elements in this specific context, due to the exclusive meaning of *only*. Similarly, if an element often co-occurs with *even* in a specific context, we might be able to infer from this that the element is often unexpected or unlikely in that context. Good models that are able to capture long-distance dependencies, should be able to then capture that these elements are only in certain contexts unexpected or not accompanied by other elements. We especially chose to use LSTM models for this task, as we think they are able to capture these dependencies. Also for this reason, we still believe in the possibility of the models to capture human intuitions on alternative sets.

4.3 Generating Alternatives

This section is divided into four parts and the aim is to investigate how well the LMs are able to capture human intuitions on the ordering of the alternatives. We start by analyzing the naturalness of sentences with correct and incorrectly hypothesized ordering. After that, we investigate how well the models are able to predict the correct ordering. Finally, we compare the ordering of the human generated alternatives through investigating the position in the list of generated alternatives by the model. Similar to the previous section, we close this section with a brief discussion.

4.3.1 Naturalness of Human Alternatives

In this section, we compare the naturalness of sentences with the human generated alternatives on different conditions to explore the ability of the models of capturing the scalar properties of *even*. We performed an experiment where we compared the naturalness, expressed in terms of SLOR and perplexity of the original sentence with the human generated alternative to the sentence where we switch the object and the alternative as in (34). In a similar fashion, wee also did an experiment where we deleted the ending of the sentences and compare the naturalness of the original object to the alternative in the object position as in (35). For convenience, these setups have been copied here in (43) and (44).

(43) a. Condition 1: *Subject Verb Object* and even *Alternative*.

b. Condition 2: *Subject Verb Alternative* and even *Object*.

(44) a. Condition 1: *Subject Verb Object*.

b. Condition 2: *Subject Verb Alternative*.

	SLOR		PPL	
	Cond. 1	**Cond. 2**	**Cond. 1**	**Cond. 2**
Switching object and alternative	1.545	1.539	297.9	297.3
Deleting object versus alternative	1.333	1.281	264.6	341.9

Table 4.12: The naturalness of different conditions on the switching task and deletion task for the LSTM model, using both SLOR and perplexity (PPL).

	SLOR		PPL	
	Cond. 1	**Cond. 2**	**Cond. 1**	**Cond. 2**
Switching object and alternative	4.960	4.784	9.5	11.9
Deleting object versus alternative	3.572	3.658	30.5	32.5

Table 4.13: The naturalness of different conditions on the switching task and deletion task for the BiLSTM model, using both SLOR and perplexity (PPL).

As mentioned before, we expect in both setups that Condition 1 is more natural in general, since in Condition 2, we either reverse the scale of *even*[23] or use the most implausible alternative in object position.

The results of the naturalness of the sentences, divided per setup and condition, are summarized in Table 4.12. In this table, the sentences are evaluated on the LSTM, using both evaluation measures. On the switching task, the model's perceived naturalness is very similar for both conditions, for both SLOR and perplexity. This means that the switching the alternatives does not seem to negatively affect the naturalness of the sentences, as this is the only factor that was different between the conditions.

Condition 1 is perceived as slightly more natural on the deletion task when the LSTM model is evaluated using both SLOR and PPL. Since the differences are small and since we did not find any differences on the switching task, we suspect that there is again no effect of ordering of the alternatives on the naturalness of the sentences. Instead, we suspect that this difference is caused by a difference in probability of the element in a given context.

The results of performing the same tasks on the BiLSTM model are summarized in Table 4.13. We see the opposite effect when comparing the results to the LSTM model. For the BiLSTM model we can see a small positive difference in naturalness between the conditions for the switching task and even a small negative on the deletion task. The evaluation on the deletion task of the BiLSTM model using perplexity on the other shows a positive difference. Again, this difference might be caused by frequency effects.

As mentioned earlier, the results could be influenced by both alternatives being plausible or scale changes. Therefore, we repeated the same experiment on a subset of all sentences where this switch is not possible or very unlikely. We manually selected

[23]We found this effect when analyzing the generated alternatives in Section 4.1.2. In this case, we don't expect to find any difference between the conditions.

	SLOR		PPL	
	LSTM	**BiLSTM**	**LSTM**	**BiLSTM**
Switching object and alternative	42.54	68.30	42.54	68.30
Deleting object versus alternative	62.54	43.49	73.97	61.90

Table 4.14: The percentage of times the model predicts Condition 1, as compared to Condition 2 by comparing the naturalness of a sentence. The evaluation is done on both the LSTM model and BiLSTM model, using both SLOR and perplexity (PPL).

these sentences, using our own intuitions. We found that sentences where switching the alternatives did not seem equally natural, still were perceived as equally plausible, suggests further that the LSTM model is not able to capture the scalar properties of *even*, using the current setup.

Overall, both models at a first glance do not seem to measure differences in naturalness, due the change in the ordering of the alternatives and therefore do not seem to capture information about the ordering of alternatives in a set of alternatives.

4.3.2 Predicting the Ordering of Alternatives

We investigated how often Condition 1 is perceived as more natural by the models, compared to Condition 2 on both the switching and deletion task. A higher accuracy would indicate that the models have learned something about the ordering of the alternatives. The results are reported for both evaluation measures and both models and are summarized in Table 4.14.

First, notice that the accuracy of the models on the switching task is the same for SLOR and perplexity. This is because the sentences contain the same words and only differ in their word order. The only factor that causes differences in results between the evaluation metrics is the correction of frequency biases in SLOR through subtracting the unigram probability from the whole sentence. Both evaluation metrics use the loss of the models in some way and compensate for sentence length[24]. This means that when the same words are used, the evaluation metrics are proportional to each other. The consequence is that they make exactly the same predictions.

What we can observe from the results on switching task that the LSTM model does not seem able to capture the scale differences with an accuracy that is below chance. The BiLSTM model, on the other hand, seems to capture something about the scales. This is quite surprising, since the BiLSTM is not able to capture any human intuitions on focus particles in the focus particle prediction task.

On the deletion task, we see the opposite effect where the LSTM model seems to capture some intuitions on the most plausible alternative, in contrast to the BiLSTM model. Contrary to the results on the comparison between *even* and *only*, we now observe that the performance on perplexity is higher on the deletion task. This, as

[24]Which is equal in both conditions as well.

mentioned before, seems to be an effect of the frequency differences between the object and alternative.

If the models were able to capture differences in the likelihood or unexpectedness of alternatives, we would expect to see consistent results on both tasks. However, we do not see consistent results and therefore it seems that both models are not able to capture which alternative is more plausible in a given context and thus capture the scalar properties of *even*.

4.3.3 Ability of Generating Plausible Alternatives

In the final part of this thesis, we analyze the actual alternatives generated by the model. We only performed this analysis on the LSTM model as the BiLSTM model has not been able to capture any information about alternative sets in our previous experiments. For this experiment, we have three conditions for which we generated alternatives. The first condition is the sentence as we presented to humans. The second condition is the same, but without the word *even* and in the third condition, the alternative is in the position of the object. The three conditions are shown in (45).

(45) a. Condition 1: *Subject Verb Object* and even ___.

b. Condition 2: *Subject Verb Object* and ___.

c. Condition 3: *Subject Verb Alternative* and even ___.

For the three conditions we generated the 1000 most plausible alternatives and analyzed how many of the human generated alternatives are in these lists. We would expect that the ranks of the alternatives in Condition 1 are generally higher than for Condition 2 and 3. We expect the model to rank the implausible alternatives higher when *even* is used, as opposed to Condition 2 where *even* is not used. This would mean that there are more plausible alternatives in Condition 2 and thus that the ranks are generally lower. We also expect the ranks of the alternatives in Condition 3 to be lower, as compared to Condition 1, because of switching the alternatives.

The results are summarized in Table 4.15. Note that there are only 171 alternatives in Condition 3. That is because we compare the list of alternatives generated by the model not to the list of human generated alternatives, but to the list of objects of the sentences. This means that we have to remove the objects that consist of more than one word from the retrievable human alternatives, since they cannot be retrieved by the model.

From Table 4.15, we can observe that in Condition 2 more human generated alternatives were retrieved than in Condition 1. Also, in Condition 3 a lot more alternatives were retrieved. We would expect to retrieve more human alternatives in Condition 1 than in Condition 2 and 3, as explained before. These results are thus in contrast to our expectations.

In Table 4.16, we report the percentage of times that the rank of the alternative in Condition 1 was higher than the rank of the same alternatives in Condition 2 or

Condition	# Alternatives
Cond. 1: Normal sentence	94/209
Cond. 2: Sentence without even	115/209
Cond. 3: Sentence order changed	130/171

Table 4.15: The number of alternatives in the list of 1000 alternatives that were also generated by the human participants, per condition. The number after the slash symbol indicates the total number human alternatives there were for that condition.

Comparison Conditions	Percentage
Cond. 1 versus Cond. 2	8.70
Cond. 1 versus Cond. 3	33.33

Table 4.16: The percentage of times the alternative in Condition 1 has a higher rank than the alternative in Condition 2 in the LSTM model.

the rank of the objects in Condition 3. We find that the ranks of the alternatives in Condition 2 are almost always higher than the ranks of the alternatives in Condition 1. There could be multiple reasons for this behavior. One possible explanation for this, is that both structures might have different plausible syntactic continuations. For example, it could be that the model has seen more sentential uses of *even*. In that case the subject of the new sentence could still be noun, but is not (very) related to the explicit alternative. Therefore the model might have learned to partially associate the wrong alternatives with *even*. Besides, the position of the alternatives in Condition 1 is further away from the verb and the subject than the position of the alternatives in Condition 2. Because the LSTM model processes the sentence incrementally and the previous word has the largest influence on the prediction of the current word. This means that the information on the subject and verb is less used in the prediction. Finally, there is much more data available for the possible alternatives in Condition 2. This means it might not have seen enough relevant alternatives and it might also not have formed any associations with the human generated alternatives.

We can furthermore observe from Table 4.16 that the ranks of the alternatives in Condition 3 are generally higher than the ranks of the alternatives in Condition 1. The alternatives provided by humans are probably less frequent, because they are less expected. This seems to suggest that the frequency of the alternatives has a large influence on the rank. We also will note again that in many cases the object and alternative can be switched without affecting the naturalness of the sentence. This might also play a role in the results. This again indicates that the LSTM model does not seem able to capture the scalar properties of *even*.

Overall, we have not found any evidence that the alternatives that the LSTM model generates are restricted or influenced by *even*. This would have been an indication that the model is able to capture some human intuitions on alternative sets. Since there are many things that could have resulted in the negative results, besides the inability

of the models to learn something about the human intuitions on alternatives, we are not able to draw definitive conclusions on how good the LSTM model is at capturing human intuitions on alternatives.

4.3.4 Discussion

Considering the results on comparing the generated alternatives, we do not get consistent results about any of the models being able to capture any effects about the ordering of the generated alternatives. Both models seem able to capture some intuitions on the scalar properties of *even* when they are evaluated using perplexity. However, these effects seem largely due to the frequencies of the alternatives and do not seem due to the scalar effects of *even*. The LSTM model also does not seem able to capture the scalar effects of *even* in the ranks of the alternatives. Combining the results of both parts, we have to conclude that the models do not seem able to capture any scalar effects of *even*, and, in turn, do not seem to capture the human intuitions on the scalar effects of *even*.

This does not necessarily mean they are not able to capture these effect at all. There could be multiple reasons why we do not observe any effects. First, even with the restrictions put on the alternatives through focus particles, the set of alternatives is still very large and diverse. We might simply not have found a good way to filter the list of alternatives. Besides, the human participants did not provide us with the alternatives we anticipated for, which might have been caused by our sentences not being the right stimuli. Also, our experimental setup might not have been correct. The consequence is that the analyses are not completely valid, because in some cases switching the alternatives does not lead to a less natural sentence. This, in turn, might be a reason for why we did not find consistent results that provided evidence that the models capture human intuitions on *even* and the ordering of alternatives. The easiness with which the scale of *even* changes in the experiment, might also be due to the flexibility of these scales. Even when considering the contexts, the scales could still be easily reversed, which could point toward the end of the scale being lexically unspecified and only determined by the context, as argued by Rullmann (2007).

Chapter 5

Conclusion

The goal of this thesis was to explore the possibilities of modeling alternatives with language models, in particular with RNNs. We used focus particles to restrict the possible set of alternatives and investigated how they affect this set of alternatives as they put quite strong restrictions on the set of alternatives and this is their only function.

In the first experiment, we collected a corpus of simple natural sentences with focus particles *even* and *only*. We left out the focus particle and asked human participants to indicate the most plausible focus particle in these sentences. Using these judgments as gold standard, we evaluated two neural language models (LMs), using two different evaluation metrics on these judgments to investigate to what extent the models are able capture the human intuitions on focus particles. We used a simple LSTM model and a simple BiLSTM model, because they have proven to be successful in different language modeling tasks, including grammaticality judgments. We evaluated both models using SLOR, a relatively new and promising evaluation metric, and perplexity, the most-used evaluation metric for LMs.

We found three pieces of evidence pointing toward the LSTM model being able to capture the human intuitions on the focus particles *even* and *only*. First, the sentences with the correct focus particle, as judged by the human participants, were seen as more natural by the model. Second, the exact meaning contribution of the correct focus particle was significantly larger than the meaning contribution of the incorrect focus particle, suggesting that the focus particles have learned the correct semantics of the focus particles. Finally, the LSTM model was able to correctly predict the most plausible focus particle in many cases. These consistent results and the fact the isolated contribution of the correct versus incorrect focus particle was significant, is strong evidence that the LSTM model is able to capture differences in alternative sets, induced by focus particles.

The BiLSTM model, in contrast, did not seem able to capture the human intuitions on focus particles. We have argued that this could be due to a ceiling effect where the model is not able to distinguish between the subtle differences in the alternatives sets, induced by the different focus particles, because it is too optimized for measuring the

predictability of a sentence. Since the sentences are, in fact, all grammatical, the sentences might be seen as equally natural. The role of the evaluation metrics might play a role as well. We have seen that both SLOR and perplexity are sensitive to frequency biases in a specific syntactic position. Especially perplexity, and SLOR to some extent as well, seem to rely too much on syntactic information in deciding the naturalness of the sentences. Therefore it seems that perplexity, and SLOR to a lesser a extent, are not very suitable for making semantic judgments. Because the LMs rely strongly on word order, we are also not sure if there are evaluation metrics for them where we can solely evaluate semantic behavior.

We also investigated the influence of the syntactic position on the human and models' judgments of focus particles. We found that for both the models and the human participants, the object position seemed the easiest position to judge. We found three potential explanations. Since humans process sentences incrementally, they might have the most information available in object position, which is similar to the LSTM model that also processes the sentence incrementally. Besides, computing alternatives seems harder for verbs than for nouns and there is ambiguity if the focus particle is in verb position, because it could associate with both the verb and the object.

In the second part of this thesis, we investigated the ability of LMs to capture the scalar properties of *even*, by looking at the human generated alternatives and the alternatives, generated by the LSTM model. First, we manually contructed a small corpus of sentences where the humans and models had to generate an alternative, given a sentence, an explicit alternative and the focus particle *even* that marks that this alternative is more unxpected than the explicit alternative. We investigated the influence of the ordering of the alternatives on the naturalness of the sentence and also investigated the influence of *even* on the ordering of the alternatives in the LSTM model. From our findings, we did not find any consistent evidence that the models are able to capture the scalar effects of *even*. We also did not find any evidence that the ranks of the alternatives in the LSTM model are influenced by the restrictions put on them through *even*.

There are still several problems with our approach on generating focus particles. The alternatives that the participants generated often do not seem extreme enough or more unexpected than the object that it is compared to. Therefore it is not very surprising that the LMs also were not able to capture our expectations on the scalar properties of *even*. In a future experimental design, it might be a good idea to provide more context and restrict the set of possible alternatives more, to avoid the lack of context to result in the possibility of reversing the scale of *even* without any problems. The scales for *even* also seem more flexible than we expected, which could be the result of this lack of context. It could however also be a property of the *even*, which would be an argument in favor of the theory by Rullmann (2007), that states that the end of the scale of *even* is lexically unspecified and is determined by a salient property in the context.

This thesis is a first step in using LMs for generating alternatives. The LSTM model seems able to learn something about alternative sets, although this was maybe

not as convincing as we had hoped. More research is needed to explore this potential and the possible limitations. The models were not optimized for the best performance. However, we have to be careful not to filter out the subtle differences in naturallness, induced by differences in alternative sets in the optimization. As we have seen, the more optimized BiLSTM model did not perform better and seemed to have lost the ability to detect these subtle differences. Therefore we probably have to apply a significant amount of regularization to prevent the models from becoming too optimized. We are also curious how more sophisticated and more complex state-of-the-art LMs, such as BERT (Devlin et al., 2018) and the model by Radford et al. (2019) would perform on our experiments.

It would also be interesting to explore the ability of LMs to capture other focus particles such as *just* and *also*. Finally, we could study the alternatives induced by focus particles in more complex sentences and perhaps even study which element the focus particle associates with, in more ambiguous cases. Finally, we could investigate the nature of the scales, induced by some focus particles more systematically, as, for example, has been done in Distributional Semantics (DS) for adjectives (Aina, 2017). Another direction is to explore the ability of LMs to capture other alternatives-inducing phenomena, such as negation, quantifiers such as *many* and *most* and generic sentences. Because these phenomena, in contrast to focus, do not rely on intonational cues, they might be easier to model using LMs. Besides, alternatives induced by negation have been successfully modeled using DS techniques (Kruszewski et al., 2016). The directions outlined here could help us in making the alternatives more explicit and give us insight in these linguistic phenomena, especially on the syntactic side of them, we expect, because the current LMs perform best in modeling syntactic information.

Bibliography

Abel, A. D., Maguire, M. J., Naqvi, F. M., and Kim, A. Y. (2015). Lexical retrieval of nouns and verbs in a sentence completion task. *Journal of psycholinguistic research*, 44(5):545–553.

Aina, L. (2017). Not logical: A distributional semantic account of negated adjectives. Master's thesis, Universiteit van Amsterdam.

Arisoy, E., Sainath, T. N., Kingsbury, B., and Ramabhadran, B. (2012). Deep neural network language models. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pages 20–28. Association for Computational Linguistics.

Atlas, J. D. (1993). The importance of being 'only': Testing the neo-gricean versus neo-entailment paradigms. *Journal of Semantics*, 10(4):301–318.

Bayer, J. (1996). *Directionality and Logical Form: On the Scope of Focusing Particles and Wh-in-situ*. Kluwer Academic Publishers.

Beaver, D. I. and Clark, B. Z. (2008). *Sense and Sensitivity: How Focus Determines Meaning*. Wiley–Blackwell.

Benz, A. (2010). Optimal completion and implicatures of complex sentences: a game theoretic approach. In *Proceedings of the 7th workshop of logic and engineering of natural language semantics (LENLS)*, pages 41–55.

Bernardy, J.-P. and Lappin, S. (2017). Using deep neural networks to learn syntactic agreement. *Linguistic Issues in Language Technology LiLT*, 15:1–15.

Bott, L. and Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of memory and language*, 51(3):437–457.

Bowman, S. R., Manning, C. D., and Potts, C. (2015). Tree-structured composition in neural networks without tree-structured architectures. In *Proceedings of the 2015th International Conference on Cognitive Computation: Integrating Neural and Symbolic Approaches*, volume 1583, pages 37–42.

Breedin, S. D., Saffran, E. M., and Schwartz, M. F. (1998). Semantic factors in verb retrieval: An effect of complexity. *Brain and language*, 63(1):1–31.

Buccola, B., Križ, M., and Chemla, E. (2018). Conceptual alternatives: Competition in language and beyond.

Büring, D. (2016). Unalternative semantics. In *Semantics and Linguistic Theory*, volume 25, pages 550–575.

Byram-Washburn, M. (2013). *Narrowing the Focus: Experimental studies on exhaustivity and contrast*. PhD thesis, University of Southern California.

Charniak, E. et al. (2016). Parsing as language modeling. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2331–2336.

Chemla, E. and Singh, R. (2014). Remarks on the experimental turn in the study of scalar implicature. *Language and Linguistics Compass*, 8(9):373–399.

Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. *Structures and beyond*, 3:39–103.

Chierchia, G. (2006). Broaden your views: Implicatures of domain widening and the "logicality" of language. *Linguistic inquiry*, 37(4):535–590.

Chierchia, G. (2013). *Logic in Grammar: Polarity, Free Choice, and Intervention*, volume 2. Oxford University Press.

Chierchia, G., Fox, D., and Spector, B. (2012). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. *Semantics: An international handbook of natural language meaning*, 3:2297–2332.

Chowdhury, S. A. and Zamparelli, R. (2018). Rnn simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 133–144.

Cohen, A. (1999). How are alternatives computed? *Journal of Semantics*, 16(1):43–65.

Coppock, E. and Beaver, D. I. (2014). Principles of the exclusive muddle. *Journal of Semantics*, 31:371–432.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., and Salakhutdinov, R. (2019). Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. *arXiv preprint arXiv:1901.02860*.

Damasio, A. R. and Tranel, D. (1993). Nouns and verbs are retrieved with differently distributed neural systems. *Proceedings of the National Academy of Sciences*, 90(11):4957–4960.

Damasio, H., Grabowski, T. J., Tranel, D., Hichwa, R. D., and Damasio, A. R. (1996). A neural basis for lexical retrieval. *Nature*, 380(6574):499–505.

Degen, J. and Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive science*, 40(1):172–201.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dyer, C., Kuncoro, A., Ballesteros, M., and Smith, N. A. (2016). Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 199–209.

Erlewine, M. Y. (2014). Explaining leftward focus association with even but not only. In *Proceedings of Sinn und Bedeutung*, volume 18, pages 128–145.

Erlewine, M. Y. (2016). Focus association into copies and the scope of even. In *Semantics and Linguistic Theory*, volume 26, pages 855–873.

Erlewine, M. Y. (2018). Lecture notes on the syntax and semantics of focus particles.

Fauconnier, G. (1976). *Etude de certains aspects logiques et grammaticaux de la quantification et de l'anaphore en français et en anglais*. Champion, Paris.

Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54.

Filik, R., Paterson, K. B., and Liversedge, S. P. (2009). The influence of only and even on online semantic interpretation. *Psychonomic Bulletin & Review*, 16(4):678–683.

Fillmore, C. (1965). Entailment rules in a semantic theory. *The Ohio State University Project on Linguistic Analysis*, 10.

Fox, D. (2007). Free choice and the theory of scalar implicatures. In *Presupposition and implicature in compositional semantics*, pages 71–120. Springer.

Frank, M. C. and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Franke, M. (2011). Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics and Pragmatics*, 4:1–82.

Franke, M. and Jäger, G. (2016). Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift für sprachwissenschaft*, 35(1):3–44.

Fraundorf, S. H., Benjamin, A. S., and Watson, D. G. (2013). What happened (and what did not): Discourse constraints on encoding of plausible alternatives. *Journal of memory and language*, 69(3):196–227.

Fried, D., Stern, M., and Klein, D. (2017). Improving neural parsing by disentangling model combination and reranking effects. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 161–166.

Fălăuş, A. (2013). Introduction: Alternatives in semantics and pragmatics. In *Alternatives in semantics*, pages 1–35. Palgrave Macmillan UK, London.

Gast, V. and Rzymski, C. (2015). Towards a corpus-based analysis of evaluative scales associated with even. *Linguistik online*, 71(2).

Geurts, B. (2010). *Quantity implicatures*. Cambridge University Press.

Giannakidou, A. (2007). The landscape of even. *Natural Language & Linguistic Theory*, 25(1):39–81.

Goodman, N. D. and Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184.

Gotzner, N. (2014). What's included in the set of alternatives? psycholinguistic evidence for a permissive view. In *19th Annual Conference Sinn und Bedeutung*.

Gotzner, N., Wartenburger, I., and Spalek, K. (2016). The impact of focus particles on the recognition and rejection of contrastive alternatives. *Language and Cognition*, 8(1):59–95.

Greenberg, Y. (2015). Even, comparative likelihood and gradability. In *Proceedings of the Amsterdam colloquium*, volume 20, pages 147–156.

Greenberg, Y. (2016). A novel problem for the likelihood-based semantics of even. *Semantics and Pragmatics*, 9:1–28.

Greenberg, Y. (2018). Even and only: Arguing for parallels in scalarity and in constructing focus alternatives. *North East Linguistics Society (NELS)*, 49.

Greenberg, Y. (2019). Scalarity, exclusivity, mirativity/evaluativity: What (and what doesn't) make 'only' a mirror image of 'even'.

Grice, H. P. (1975). Logic and conversation. In Cole, P. and Morgan, J., editors, *Syntax and Semantics*, volume 3: Speech Acts, pages 41–58. Academic Press.

Grice, H. P. (1989). *Studies in the Way of Words*. Harvard University Press.

Groenendijk, J. A. G. and Stokhof, M. J. B. (1984). *Studies on the semantics of questions and the pragmatics of answers*. PhD thesis, University of Amsterdam.

Guerzoni, E. (2003). *Why even ask? On the pragmatics of questions and the semantics of answers*. PhD thesis, Massachusetts Institute of Technology.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., and Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 1195–1205.

Hamblin, C. L. (1973). Questions in montague english. *Foundations of Language*, 10(1):41–53.

Herburger, E. (2000). *What counts: Focus and quantification*, volume 36. MIT Press.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Horn, L. (1969). A presuppositional analysis of only and even. In *Proceedings of the Fifth Regional Meeting*. Chicago Linguistics Society.

Horn, L. (1989). *A Natural History of Negation*. Chicago University Press.

Horn, L. R. (1972). *On the semantic properties of logical operators in English*. PhD thesis, University of California, Los Angeles.

Horn, L. R. (1996). Exclusive company: Only and the dynamics of vertical inference. *Journal of semantics*, 13(1):1–40.

Huang, Y. T. and Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive psychology*, 58(3):376–415.

Husband, E. M. and Ferreira, F. (2016). The role of selection in the comprehension of focus alternatives. *Language, Cognition and Neuroscience*, 31(2):217–235.

Huttenlocher, J. and Lui, F. (1979). The semantic organization of some simple nouns and verbs. *Journal of verbal learning and verbal behavior*, 18(2):141–162.

Jackendoff, R. S. (1972). *Semantic interpretation in generative grammar*. MIT Press.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P., and Suchomel, V. (2013). The tenten corpus family. In *7th International Corpus Linguistics Conference CL*, pages 125–127.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. (2016). Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Kamide, Y., Altmann, G. T., and Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and language*, 49(1):133–156.

Karttunen, L. (1977). Syntax and semantics of questions. *Linguistics and Philosophy*, 1(1):3–44.

Karttunen, L. and Peters, S. (1979). Conventional implicature. In Oh, C.-K. and Dinneen, D. A., editors, *Syntax and Semantics*, volume 11, pages 1–56. Academic Press.

Katzir, R. (2013). A note on contrast. *Natural language semantics*, 21(4):333–343.

Kay, P. (1990). Even. *Linguistics and philosophy*, 13(1):59–111.

Kim, C. S. (2012). *Generating alternatives: Interpreting focus in discourse*. PhD thesis, University of Rochester.

Kim, S. (2011). *Focus particles at syntactic, semantic and pragmatic interfaces: The acquisition of only and even in English*. PhD thesis, University of Hawaii.

Klinedinst, N. (2005). Scales and only. Master's thesis, University of California, Los Angeles.

König, E. (1991). *The meaning of focus particles: A comparative perspective*. Routledge, London.

Krifka, M. (1992). A compositional semantics for multiple focus constructions. In *Informationsstruktur und grammatik*, pages 17–53. Springer.

Kruszewski, G., Paperno, D., Bernardi, R., and Baroni, M. (2016). There is no logical negation here, but there are alternatives: Modeling conversational negation with distributional semantics. *Computational Linguistics*, 42(4):637–660.

Kuncoro, A., Dyer, C., Hale, J., Yogatama, D., Clark, S., and Blunsom, P. (2018). Lstms can learn syntax-sensitive dependencies well, but modeling structure makes them better. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1426–1436.

Lahiri, U. (2008). The of some scalar expressions in spanish. *Anuario del Seminario de Filología Vasca" Julio de Urquijo"*, 42:359–389.

Lassiter, D. and Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Semantics and linguistic theory*, volume 23, pages 587–610.

Lau, J. H., Clark, A., and Lappin, S. (2016). Grammaticality, acceptability, and probability: a probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

Ling, W., Dyer, C., Black, A. W., and Trancoso, I. (2015a). Two/too simple adaptations of word2vec for syntax problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1299–1304.

Ling, W., Dyer, C., Black, A. W., Trancoso, I., Fermandez, R., Amir, S., Marujo, L., and Luis, T. (2015b). Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1520–1530. Association for Computational Linguistics.

Linzen, T., Dupoux, E., and Goldberg, Y. (2016). Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association of Computational Linguistics*, 4(1):521–535.

Luong, T., Kayser, M., and Manning, C. D. (2015). Deep neural language models for machine translation. In *Proceedings of the 19th Conference on Computational Natural Language Learning*, pages 305–309.

Masterson, J., Druks, J., and Gallienne, D. (2008). Object and action picture naming in three-and five-year-old children. *Journal of Child Language*, 35(2):373–402.

Mätzig, S., Druks, J., Masterson, J., and Vigliocco, G. (2009). Noun and verb differences in picture naming: Past studies and new evidence. *Cortex*, 45(6):738–758.

McCawley, J. D. (1981). *Everything that linguists have always wanted to know about logic... but were ashamed to ask*. University of Chicago Press.

Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *11th Annual Conference of the International Speech Communication Association*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, 78(2):165–188.

Paterson, K. B., Liversedge, S. P., Filik, R., Juhasz, B. J., White, S. J., and Rayner, K. (2007). Focus identification during sentence comprehension: Evidence from eye movements. *The Quarterly Journal of Experimental Psychology*, 60(10):1423–1445.

Pauls, A. and Klein, D. (2012). Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 959–968.

Plank, B., Søgaard, A., and Goldberg, Y. (2016). Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *The 54th Annual Meeting of the Association for Computational Linguistics*, pages 412–418.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A comprehensive grammar of the English language*. Pearson Longman.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. Technical report, OpenAI.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Roberts, C. (1996). Information structure in discourse: Towards an integrated formal theory of pragmatics. *OSU Working Papers in Linguistics*, 49:91–136.

Roberts, C. (2011). Only: A case study in projective meaning. In *Baltic International Yearbook of Cognition, Logic and Communication*, volume 6 of *Formal Semantics and Pragmatics: Discourse, Context, and Models*, pages 1–59.

Rooth, M. (1985). *Association with focus*. PhD thesis, University of Massachusetts, Amherst.

Rooth, M. (1992). A theory of focus interpretation. *Natural language semantics*, 1(1):75–116.

Rullmann, H. (1997). Even, polarity, and scope. In Martha Gibson, G. W. and Libben, G., editors, *Papers in experimental and theoretical linguistics*, volume 4, pages 40–64. Department of Linguistics, University of Alberta, Edmonton, Canada.

Rullmann, H. (2007). What does even even mean.

Rullmann, H. and Hoeksema, J. (1997). De distributie van ook maar en zelfs maar: een corpusstudie. *Nederlandse taalkunde*, 2(28).

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and philosophy*, 27(3):367–391.

Sauerland, U. (2012). The computation of scalar implicatures: Pragmatic, lexical or grammatical? *Language and Linguistics Compass*, 6(1):36–49.

Schwenk, H., Rousseau, A., and Attik, M. (2012). Large, pruned or continuous space language models on a gpu for statistical machine translation. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pages 11–19. Association for Computational Linguistics.

Storozum, J. (2018). Opposites attract - or do they?: Investigating negated verbs in distributional semantic space. Master's thesis, Brandeis University.

Szabolcsi, A. (2017). Additive presuppositions are derived through activating focus alternatives. In *Proceedings of the 21st Amsterdam colloquium*, pages 455–465.

Taglicht, J. (1984). *Message and emphasis: On focus and scope in English*. Pearson Longman.

Tanenhaus, M. K. and Brown-Schmidt, S. (2008). Language processing in the natural world. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1493):1105–1122.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

Tessler, M. H., Lopez-Brau, M., and Goodman, N. D. (2017). Warm (for winter): Comparison class understanding in vague language. In *15th International Conference on Cognitive Modeling*, page 193.

Trampuš, M. and Novak, B. (2012). Internals of an aggregated web news feed. In *Proceedings of the 15th Multiconference on Information Society 2012*, volume 12.

Traxler, M. J., Bybee, M. D., and Pickering, M. J. (1997). Influence of connectives on language comprehension: eye tracking evidence for incremental interpretation. *The Quarterly Journal of Experimental Psychology: Section A*, 50(3):481–497.

Trinh, T., Dai, A., Luong, T., and Le, Q. (2018). Learning longer-term dependencies in rnns with auxiliary losses. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4972–4981.

Van Rooij, R. and Schulz, K. (2004). Exhaustive interpretation of complex sentences. *Journal of logic, language and information*, 13(4):491–519.

Van Rooij, R. and Schulz, K. (2007). Only: Meaning and implicatures. In Aloni, M., Butler, A., and Dekker, P., editors, *Questions in dynamic semantics*, volume 17 of *Current Research in the Semantics/Pragmatics Interface*, pages 193–223. Brill.

Vigliocco, G., Vinson, D. P., Druks, J., Barber, H., and Cappa, S. F. (2011). Nouns and verbs in the brain: a review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience & Biobehavioral Reviews*, 35(3):407–426.

Von Stechow, A. (1991). Current issues in the theory of focus. In Von Stechow, A. and Wunderlich, D., editors, *Semantik: Ein Internationales Handbuch Der Zeitgenössischen Forschung*, pages 804–824. Walter de Gruyter.

Wagner, M. (2005). *Prosody and recursion*. PhD thesis, Massachusetts Institute of Technology.

Wagner, M. (2006). Givenness and locality. In *Semantics and Linguistic Theory*, volume 16, pages 295–312.

Wagner, M. (2012). Focus and givenness: a unified approach. In Kučerová, I. and Neeleman, A., editors, *Contrasts and Positions in Information Structure*, pages 102––147. Cambridge University Press.

Wagner, M. (2014). Even and the syntax of focus sensitivity. In *A paper presented at the workshop on "Focus Sensitive Expressions from a Cross-Linguistic Perspective", Bar Ilan University.*

Wagner, M. (2015). Additivity and the syntax of even. *Linguistics Colloquium, University of Chicago.*

Wilcox, E., Levy, R., Morita, T., and Futrell, R. (2018). What do rnn language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221.

Wilkinson, K. (1996). The scope of even. *Natural language semantics*, 4(3):193–215.

Winterstein, G. (2012). Only without its scales. *Sprache und Datenverarbeitung*, 35–36:29–47.

Zeevat, H. (2009). "only" as a mirative particle.

Zeevat, H. (2013). Expressing surprise by particles. In Gutzmann, D. and Gärtner, H.-M., editors, *Beyond Expressives: Explorations in Use-Conditional Meaning*, volume 28 of *Current Research in the Semantics/Pragmatics Interface*, pages 297–320. Brill.