# The Acquisition of Semantic Representations of Anaphora in Complex Expressions

Jackson Petty
jackson.petty@yale.edu

Summer 2020
Yale University
New Haven, Connecticut

Advised by Robert Frank

# Contents



# 1 Introduction

Well-formed $\bullet\!\!-\!\!\circ$ constructions of a natural language have both an observable representation, written or spoken words, and a semantic representation, the meaning behind these utterances. Taken together, the observable and semantic represenations of a phrase constitue a form-meaning pair. Knowledge of that language is then the ability to translate between form and meaning, interpreting the observable form of an construction and producing new observable representations in kind from semantic representations. While to us the connection between a form and its meaning is intuitive, to machines this relationship is not innate and must then be taught.

One challenge in teaching machines to understand language comes from words or phrases whose meaning is necessarily context-dependent. Simple words, like proper nouns or verbs, may have meanings which are derivable from the words themselves. For instance, consider a simple language consiting of names and intransitive verbs like the one shown below in (1).

(1) a. $[\![\text{Alice}]\!]$ = the girl named Alice
b. $[\![\text{Bob}]\!]$ = the boy named Bob
c. $[\![\text{thinks}]\!]$ = the act of thinking

The meanings of sentences in this simple language are nicely composed of the meanings of the constituent words. For example, in (2) we see that the sentence "Alice thinks" is understood of one knows the meanings of "Alice," "thinks," and the rules about how sentences are formed from constituent parts.

(2) $[\![\text{Alice thinks}]\!] \approx [\![\text{Alice}]\!] + [\![\text{thinks}]\!]$ = the girl named Alice thinks

Not all words, however, have this nice property that their meanings are independent of the words surrounding them. Consider the case of anaphora, where reflexive pronouns refer to other nouns in a sentence, like in (3).

(3) a. Alice sees herself
    b. Claire sees herself

The sentences here can't be easily interpreted from the meanings of the individual words in the same way that (2) can. For one, it's clear that the word "herself" doesn't have a meaning independent of "Alice" or "Claire." Furthermore, it's apparent that the meaning of "herself" changes depending on the context. These are in fact the defining properties of reflexive pronouns: they must be bound by some independently-defined noun in order to have meaning. Cases like this pose challenges for machines being trained to translate forms into meaning.

Another problem facing linguists and computer scientists is trying to figure out what exactly is happening when machines "learn" a task. Defined extentially, one might think that a machine has learned a task when it can successfully complete the problems set before it. That is, if we give a machine a list of sentences and ask it to translate those sentences into semantic representations, and it does so correctly, we might think that the machine has learned how to interpret language.

# 2 Background

## 2.1 Form-Meaning Pairs

## 2.2 Anaphora and Binding Theory

## 2.3 Neural Networks and Anaphora

The question of whether or not neural networks can learn the semantic representations of anaphora is not new. [1] posed this problem and investigated whether or not a real-time RNN network architecture could predict the meanings of anaphors. Their results found that real-time networks were unable to successfully.

# 3 The Seq2Seq Network Architecture

## 3.1 Specifying the Form and Meaning domains

In order to generate training data consisting of form-meaning pairs, I use a Featural Context-Free Grammar to proceduraly generate sentences and then parse them into semantic representations using the `nltk` Python library. The Form domain is then all possible sentences generated by the grammar, while the meaning domain is the corresponding collection of semantic representations.

An example, minimal grammar is given below along with the sentences is generates and their associated semantic representations.

```
# file: grammar.fcfg
% start S

# Grammatical Rules
S[SEM = <?pred(?subj)>] -> NP[SEM = ?subj] VP[SEM = ?pred]
VP[SEM = <?v(?obj)>] -> V[SEM = ?v] NP[SEM = ?obj]

# Lexical Rules
NP[SEM = <\P.P(alice)>] -> Alice
VP[SEM = <\x.know(x)>] -> knows
```

# 4 Experiment: Generalizing anaphora to new antecedents

One of the simplest types of sentences involving reflexive pronouns are sentences containg only names, transitive verbs, and reflexive pronouns. An example of one such sentence is shown below in (4).

(4) Alice sees herself.

We define a simple predicate logic where verbs are mapped to predicates, and subjects and objects are mapped to the arguments to those predicates, as in (5) below.

(5) Bob sees Alice → see(bob, alice)

In cases where the object of the verb is a reflexive pronoun, like in (4), the subject and object arguments to the predicate are simply the subject of the sentence, as in (6).

(6) Alice sees herself → see(alice, alice)

Intransitive sentences are mapped to predicates with a single argument, as in (7).

(7) Bob sleeps → sleep(bob)

In order to test whether ot not a Seq2Seq model can be trained to generalize knowledge of reflexive sentences, we will selectively withhold certain sentences generated by the language grammar from the training set and see how the network performs on these novel cases. In this set of experiments, we will examine whether the Seq2Seq models can learn to parse 'Alice-reflexive' sentences like in 4, which are of the form of (8) below.

(8) Alice *verbs* herself

Our Seq2Seq model is made of recurrent encoders and decoders and can optionally implement attention. We conduct these tests using SRN, GRU, and LSTM architectures with No Attention, Additive Attention, and Multiplicative attention to gauge the effect that network architecture has on our models' abilities to learn to interpret reflexive pronouns and generalize this knowledge to new cases. For each combination of recurrent unit and attention, we train three models separately to see if there is any variance in the networks' abilities.

(9)

| **Grammar** | | |
|---|---|---|
| S | → | Name VP |
| VP | → | $V_i$ \| $V_t$ Name \| $V_t$ Refl |
| | | |
| Name | → | Alice \| Bob \| … \| Zelda |
| Refl | → | himself \| herself |
| $V_i$ | → | walks \| sleeps \| eats \| runs \| sings \| dances \| flies \| slumbers |
| $V_t$ | → | sees \| meets \| likes \| dislikes \| throws \| notices \| knows |

**Table 4.1** Context-free grammar used for Alice-* experiments.

The sentences produced by this grammar are matched to semantic representations in the following way.

| | | | | | |
|---|---|---|---|---|---|
| $x$ $V_i$ | → | $verb(x)$ | Alice sleeps | → | sleep(alice) |
| $x$ $V_t$ $y$ | → | $verb(x, y)$ | Bob knows Claire | → | know(bob, claire) |
| $x$ $V_t$ Refl | → | $verb(x, x)$ | Zelda sees herself | → | see(zelda, zelda) |

**Table 4.2** Semantic representations of generated sentences for Alice-* experiments.

For example, examples in (10) show how various types of sentences produced by the grammar are parsed into form-meaning pairs.

(10) a. *P verbs* → *verb*(*P*) [intransitive]
b. *P verbs Q* → *verb*(*P*, *Q*) [transitive]
c. *P verbs* ( himself | herself ) → *verb*(*P*, *P*) [reflexive]

Note that in (10b) the names *P* and *Q* need not be distinct. This grammar produces sentences like (11a), where the subject and the object of the sentence are the same name. As shown by (11b), these sentences have identical semantic representations to reflexive sentences.

(11) a. Alice sees Alice → see(Alice, Alice)
b. Alice sees herself → see(Alice, Alice)

## 4.1 Alice-$\alpha$: Can Alice know herself?

The simplest way to test generalizability of networks to the task of interpreting anaphoric expressions is to withold from training all reflexive sentences with a certain antecedent. In Alice-$\alpha$, we do exactly this, withholding all sentences of the form

(12) Alice *verbs* herself → *verb*(Alice, Alice) [withheld in Alice-$\alpha$]

from the train-test-val split. We then test the networks' abilities to interpret sentences where "Alice" is the antecedent of the reflexive pronoun "herself." All other sentences generated by the grammar, including transitive sentences where "Alice" appears as both the subject and object like in (13a), intransitive sentences where "Alice" is the subject like in (13b), sentences where "Alice" appears only as the object like in (13c), and sentences where "Alice" appears only as the subject like in (13d) where $P \neq$ "Alice", are included in the training, validation, and testing data.

(13) a. Alice *verbs* Alice → *verb*(Alice, Alice) [present in Alice-$\alpha$]
b. Alice *verbs* → *verb*(Alice) —
c. *P verbs* Alice → *verb*(*P*, Alice) —
d. Alice *verbs P* → *verb*(Alice, *P*) —

All sentences generated by the grammar not involving "Alice," such as those types enumerated in (14) where $P, Q \neq$ "Alice", are also present in the training, validation, and testing data.

(14) a. *P verbs* → *verb*(*P*) [present in Alice-$\alpha$]
b. *P verbs Q* → *verb*(*P*, *Q*) —
c. *P verbs* ( himself | herself ) → *verb*(*P*, *P*) —

**Summary of findings:** All networks are capable of successfully generalizing to the novel "Alice" antecedent with near-perfect accuracy. We performed 10 runs each for every combination of model architecture (`SRN`, `LSTM`, `GRU`, and `Transformer`) and attention (`None` or `Multiplicative`) and observed that all model structures were able to learn the correct generaliztaion with no degredation in performance on the test split or the generalization split (=near 100% accuracy averaged across all 10 runs.) This is notable as it shows how the Seq2Seq model structure is advantages in the task when compared to the original language

modeling approach of [??]. In these experiemnts, the combined encoding and decoding of the input sequences into the target predicate grammar statements allowed even the simplest networks (SRNs with no attention) to complete the task. In the language-modeling version of the problem, these same networks were unable to successfully identify the antecedents of reflexive pronouns in novel constructions.

## 4.2 Alice-$\beta$: But wait, doesn't Alice know Alice?

To be successful at the task presented in Alice-$\alpha$, networks need to be capable of lexical generalization in a new input: the token "Alice" does not appear as the antecedent of "herself" in training data and the network must learn to interpret exactly these types of sentences. However, models trained on the Alice-$\alpha$ do have one potentially useful piece of information at their disposal: although not exposed to sentences like "Alice *verbs* herself" in training, they are exposed to the closely-related sentences of the form "Alice *verbs* Alice." Though these sentences do not involve reflexive pronouns, their semantic representations are identical to reflexive sentences. This means that models in Alice-$\alpha$ may be biased in favor of producing the correct semantic parses since they already have knowledge that there exist sentences of the form of (13a) whose representation is the target output for novel constructions like (12).

To account for this possibility and increase the difficulty of the task set before the network, the Alice-$\beta$ experiment further withholds sentences like those in (15b) below, where "Alice" is both the subject and the object, along with the Alice-reflexive (15a) sentences withheld in Alice-$\alpha$.

(15) a. Alice *verbs* herself $\rightarrow$ *verb*(Alice, Alice) [withheld in Alice-$\beta$]
b. Alice *verbs* Alice $\rightarrow$ *verb*(Alice, Alice) —

The remaining types of sentences involving "Alice," intransitive sentences like those in (16a), transitive sentences where "Alice" is the subject but not the object like those in (16b), and transitive sentences where "Alice" is the object but not the subject like those in (16c), are all present in the training, validation, and testing datasets.

(16) a. Alice *verbs* $\rightarrow$ *verb*(Alice) [present in Alice-$\beta$]
b. Alice *verbs* $P$ $\rightarrow$ *verb*(Alice, $P$) —
c. $P$ *verbs* Alice $\rightarrow$ *verb*($P$, Alice) —

Finally, all sentences involving subjects and objects other than "Alice" are also present in the training, validation, and testing dataset, as shown below in (17) where $P, Q \neq$ Alice.

(17) a. $P$ *verbs* $\rightarrow$ *verb*($P$) [present in Alice-$\beta$]
b. $P$ *verbs* $Q$ *verb*($P$, $Q$) —
c. $P$ *verbs* ( himself | herself ) $\rightarrow$ *verb*($P$, $P$) —

To successfully generalize knowledge of (16) and (17) to interpret the Alice-reflexive sentences of (15b), a model must not only contend with new inputs but must also successfully produce a new semantic representation not previously encountered in training.

**Summary of findings:** Similarly to the Alice-$\alpha$ case. we found that all network structures were able to successfully learn the correct generalization across 10 separate runs, achieving near-perfect accuracy in all cases without regard to architecture or attention. This experiment shows how the networks can successfully learn to produce novel outputs when presented with novel inputs (i.e., the network is able to produce the experssion

(18) `VERB ( alice , alice )`

which it has never been trained on when encountering an expression like that of 12.)

### 4.3 Alice-$\gamma$: What if nobody knows Alice?

Given the networks' relatively good performance on the harder task of Alice-$\beta$, it seems that Seq2Seq models are capable of lexical generalization to novel antecedents. To explore the models' limits to this kind of generalization Alice-$\gamma$ further restricts the kinds of sentences present in the training data by withholding all sentences like (19c), where "Alice" appears as the object, in addition to the (19ab) sentences withheld in Alice-$\beta$.

(19) a. Alice *verbs* herself → *verb*(Alice, Alice) [withheld in Alice-$\gamma$]
b. Alice *verbs* Alice → *verb*(Alice, Alice) —
c. *P verbs* Alice → *verb*(*P*, Alice) —

Non-reflexive sentences where "Alice" serves only as the subject, like the intransitive sentences of (20a) or transitive sentences of (20b) where $P \neq$ "Alice", are included in the training, validation, and testing datasets.

(20) a. Alice *verbs* → *verb*(Alice) [present in Alice-$\gamma$]
b. Alice *verbs P* → *verb*(Alice, *P*) —

Additionally, sentences of the form of (21) for $P, Q \neq$ "Alice", where "Alice" does not appear at all, are likewise included in the training, validation, and testing sets.

(21) a. *P verbs* → *verb*(*P*) [present in Alice-$\gamma$]
b. *P verbs Q verb*(*P*, *Q*) —
c. *P verbs* ( himself | herself ) → *verb*(*P*, *P*) —

**SUmmary of findings:** The Alice-$\gamma$ case proved much more difficult for networks that the -$\alpha$ and -$\beta$ variants. The absence of training data where "Alice" served as tbe object of a transitive verb severely limited networks' performance on the Alice-reflexive test cases. In general, `SRN`s without attention, `GRU`s without attention, and `LSTM`s without attention all completely fail at this task, achieving essentially 0 accuracy on the generalization set. `GRU`s with multiplicative attention and `LSTM`s with multiplicative attention achieve middling performances of accuracies between 10% and 60% on the generalization set with most models clustered at the low end of that range. The only models which proved adept at learning this generalization were `SRN`s with multiplicative attention which achieved near-perfect accuracy on the generalization set. This finding is significant for two reasons. First, this experiment observes that attention has a noticable impact on model performance for this dataset: models without attention achieve essentially 0% accuracy, while models with attention do significantly better. Second, this experiment demonstrates that more complicated model structures do not necessarily achieve better performance that simpler structures. `SRN`s with attention greatly outperform all other model structures, even once which are more complicated in architecture, even those implementing attention. This finding is especially notable when compared to the results of the Alice-$\epsilon$ and Alice-$\zeta$ runs, where we observe `GRU`s and `LSTM`s achieving far-superior performance when compared to other architectures, indicating that the experimental domain greatly affects how useful attention and model architecture are: it is not simply a matter of throwing more complicated structures at more difficult problems.

## 4.4 Alice-$\gamma*$: What if nobody knows Alice?

In Alice-$\gamma$, we distinguished sentences where "Alice" was the subject of a sentence from those were "Alice" was the object of a sentence. In the input domain, this distinction is well-defined since in these simple sentences the subject always precedes the verb and the object always follows is. In the target domain of predicate logic, however, this distinction is muddled. While there is a distinct linear ordering to the arguments of a verb in the predicate logic domain, as shown below in (22), the presence of intransitive sentences makes the distinction more complicated.

(22) a. Alice *verbs* Bob → *verb*(Alice, Bob)
b. Bob *verbs* Alice → *verb*(Bob, Alice)

In intransitive sentences, there is only a single argument to the verb in the predicate logic representation. Here, we lose the correspondence bewtween subject/object positions in the input domain and subject/object positions in the target domain. While it is clear to speakers of English that such intransitive representations have a subject a no object, there is no structural reason the network has to not thing that intransitive sentences with only an object could exist, as shown in (23).[1]

(23) a. Alice *verbs* → *verb*(Alice)
b. * *Verbs* alice → *verb*(Alice)

In light of this, it is more proper to think about "first" and "second" position in the predicate logic domain rather than "subject" and "object" position.

This clarification raises issue that in Alice-$\gamma$, the network isn't baised against interpreting Alice-intransitive sentences as being sentences where Alice is the object. To clear this up, we run an additional experiment Alice-$\gamma*$ which additionally witholds all sentences were Alice is the subject of an intransitive sentence along with the Alice-object sentences originally held in Alice-$\gamma$.

(24) a. Alice *verbs* herself → *verb*(Alice, Alice) [withheld in Alice-$\gamma*$]
b. Alice *verbs* Alice → *verb*(Alice, Alice) —
c. *P verbs* Alice → *verb*(*P*, Alice) —
d. Alice *verbs* → *verb*(Alice) —

Non-reflexive sentences where "Alice" serves only as the subject, like the intransitive sentences of (25a) or transitive sentences of (25b) where $P \neq$ "Alice", are included in the training, validation, and testing datasets.

(25) a. Alice *verbs P* → *verb*(Alice, *P*) [present in Alice-$\gamma*$]

Additionally, sentences of the form of (26) for $P, Q \neq$ "Alice", where "Alice" does not appear at all, are likewise included in the training, validation, and testing sets.

(26) a. *P verbs* → *verb*(*P*) [present in Alice-$\gamma*$]
b. *P verbs Q verb*(*P*, *Q*) —
c. *P verbs* ( himself | herself ) → *verb*(*P*, *P*) —

**Summary of findings:** In general, we did not observe any major differences in performancy or learning rate between the Alice-$\gamma$ and Alice-$\gamma*$ experiments. The best accuracy of the SRNs with attention was reduced to just below 100%, and the best accuracies of the

[1] Mention unaccusative/passives here?

LSTMs were reduced to around 50%, but the relative performance of model structures did not change.

## 4.5 Alice-$\delta$: What if Alice doesn't know anyone?

In the same vein of further restricting the distribution of "Alice" in the training dataset, we also investigate what happens when the network is never exposed to sentences where "Alice" serves as the subject. This means withholding both Alice-reflexive (27a) and Alice-Alice (27b) sentences, along with transitive sentences like (27c) where $P \neq$ "Alice".

(27) a. Alice *verbs* herself $\rightarrow$ *verb*(Alice, Alice) [withheld in Alice-$\delta$]
b. Alice *verbs* Alice $\rightarrow$ *verb*(Alice, Alice) —
c. Alice *verbs* $P \rightarrow$ *verb*(Alice, $P$) —

Sentences where "Alice" is only the object, like (28a) for $P \neq$ "Alice", and Alice-intransitive sentences like (28b) are present in the training, validation, and testing data, as are all sentences (28c–e) not involving "Alice."

(28) a. $P$ *verbs* Alice $\rightarrow$ *verb*($P$, Alice) [present in Alice-$\delta$]
b. Alice *verbs* $\rightarrow$ *verb*(Alice) —
c. $P$ *verbs* $\rightarrow$ *verb*($P$) —
d. $P$ *verbs* $Q \rightarrow$ *verb*($P$, $Q$) —
e. $P$ *verbs* ( himself | herself ) $\rightarrow$ *verb*($P$, $P$) —

**Summary of findings:** Similarly to the Alice-$\gamma$ experiments, we observe that the presence of attention has a big influence on the performance of models. Those without (multiplicative) attention achieved near-0% accuracy on the generalization sets, while those with attention achieved far better results. Variance on the Alice-$\delta$ models was higher than on the -$\gamma$ varieties. Again, the highest performing structure was SRNs with attention, which mostly achieved between 80% and 100% accuracy, with one outlier attaining only 40%. GRUs with attention did, in general, poorly, with most models attaining less that 40% accuracy, though one outlier did achieve 100% accuracy. LSTMs with attention had much the same middling performance as in the -$\gamma$ cases.

## 4.6 Alice-$\delta$*

Similar to the Alice-$\gamma$* case, the question of the impact of intransitive sentences is worth considering here as well. "Withholding of all instances of 'Alice' in subject position" could seem to the models as withholding all instances of Alice which correspond to the first argument position in the predicate domain, and since grammatically subjects of intransitive sentences serve the same role as subjects of transitive ones it is natural to consider the affect that Alice-intransitive sentences have on the models' abilities to generalize. To that end, the Alice-$\gamma$* case further restricts the training set by with holding the Alice-intransitive sentences of (29d) alongisde the Alice-subject sentences of (29a–c) (again for $P \neq$ "Alice").

(29) a. Alice *verbs* herself $\rightarrow$ *verb*(Alice, Alice) [withheld in Alice-$\delta$]
b. Alice *verbs* Alice $\rightarrow$ *verb*(Alice, Alice) —
c. Alice *verbs* $P \rightarrow$ *verb*(Alice, $P$) —
d. Alice *verbs* $\rightarrow$ *verb*(Alice) —

Sentences where "Alice" is only the object, like (30a) for $P \neq$ "Alice", are present in the training, validation, and testing data, as are all sentences (30b–d) not involving "Alice."

(30) a. *P verbs* Alice → *verb*(*P*, Alice) [present in Alice-$\delta$]
b. *P verbs* → *verb*(*P*) —
c. *P verbs Q* → *verb*(*P*, *Q*) —
d. *P verbs* ( himself | herself ) → *verb*(*P*, *P*) —

**Summary of findings:** Distinct from the similarity between the -$\gamma$ and -$\gamma^*$ runs, witholding intransitive sentences had a large and unusual impact on the -$\delta$ runs. Performance of models without attention remained the same (=near 0% accuracy). Performance of SRNs with attention dropped dramatically, with most models attaining less that 20% accuracy, excepting one outlier which achieved 100% accuracy. GRU-attention performance likewise flipped, with almost all models scoring above 80% accuracy on the generalization set, excepting one poorly performing model which achieved only around 10% accuracy. In essence, performance of GRUs with attention and SRNs with attention flipped. The variance of the LSTMs with attention was increased: most models did slightly worse, achieving around 15–20% accuracy, though one model did eek out around 70% accuracy. These results are surprising for two reasons: First, it is notable to see how small changes in the dataset can have such a large impact on model performance. In contrast to the -$\gamma$/-$\gamma^*$ cases, witholding Alice-intransitive sentneces had a large effect on how well models were able to generalize. Second, it is very strange that increasing the difficulty of the training set by witholding more data actually increased the performance of GRUs with attention.

## 4.7 Alice-$\varepsilon$: Who's Alice and who's Claire?

The previous Alice-* experiments have thus far dealt with the task of lexical generalization of a single novel antecedent; we withhold some combinations of sentences containing the token "Alice" and test to see whether or not the network can learn to interpret sentences where "Alice" is the antecedent of a reflexive pronoun. While the results Alice-$\beta$ experiment were successful in this regard, there is a concern that the networks may have been able to interpret Alice-reflexive sentences in a negative sense: that is, by learning the interpretations of all other *Person*-reflexive sentences through direct example and then "filling in the blank" with "Alice" when confronted with a *Person*-reflexive combination which it has not yet been taught. When dealing with only a single novel antecedent, this outcome would look the same as if the network had truly acquired the generalization of (31).

(31) *P verbs* ( himself | herself ) → *verb*(*P*, *P*)

To determine if these networks are truly capable of the latter inference, we extend the Alice-$\beta$ experiment by withholding progressively more sentences of the form

(32) *P verbs* ( *P* | himself | herself )

and explore the networks' abilities to generalize knowledge of the interpretation of reflexive pronouns to new antecedents.

In withholding more than just a single *Person*-reflexive sentence, we must consider the fact that while speakers of English know that "himself" and "herself" have the same meaning, distinguished only by $\varphi$-features, the networks here have no such awareness. There is no *a-priori* reason for a network to associate "himself" with "herself" aside from their similarity in positional distribution in the training data. Because of this, withholding both "Alice *verbs* ( Alice | herself )" and "Bob *verbs* ( Bob | himself )" may not be qualitatively different for the network than withholding only "Alice *verbs* ( Alice | herself )" (with respect to the network's performance at interpreting the latter). Therefore, we will begin by removing progressively more *Person*-reflexive sentences with feminine antecedents.

**Alice-$\varepsilon$-2:** We withhold *Person*-reflexive sentences where "Alice" and "Claire" are the antecedents, shown below in (33).

(33) a. Alice *verbs* ( Alice | herself ) [withheld in Alice-$\varepsilon$-2]
b. Claire *verbs* ( Claire | herself ) —

The training set then included all *Person*-reflexive sentences with masculine antecedents and all remaining *Person*-reflexive sentences with feminine antecedents.

**Alice-$\varepsilon$-3:** We withhold *Person*-reflexive sentences where "Alice", "Claire", and "Eliza" are the antecedents, shown below in (34).

(34) a. Alice *verbs* ( Alice | herself ) [withheld in Alice-$\varepsilon$-3]
b. Claire *verbs* ( Claire | herself ) —
c. Eliza *verbs* ( Eliza | herself ) —

The training set then included all *Person*-reflexive sentences with masculine antecedents and all remaining *Person*-reflexive sentences with feminine antecedents.

**Alice-$\varepsilon$-6:** We withhold *Person*-reflexive sentences where "Alice", "Claire", "Eliza", "Grace", "Isla", and "Katherine" are the antecedents, shown below in (35).

(35) a. Alice *verbs* ( Alice | herself ) [withheld in Alice-$\varepsilon$-6]
b. Claire *verbs* ( Claire | herself ) —
c. Eliza *verbs* ( Eliza | herself ) —
d. Grace *verbs* ( Grace | herself ) —
e. Isla *verbs* ( Isla | herself ) —
f. Katherine *verbs* ( Katherine | herself ) —

The training set then included all *Person*-reflexive sentences with masculine antecedents and all remaining *Person*-reflexive sentences with feminine antecedents.

**Alice-$\varepsilon$-14:** We withhold *Person*-reflexive sentences where "Alice", "Claire", "Eliza", "Grace", "Isla", and "Katherine" are the antecedents, shown below in (36).

(36) a. Alice *verbs* ( Alice | herself ) [withheld in Alice-$\varepsilon$-14]
b. Claire *verbs* ( Claire | herself ) —
c. Eliza *verbs* ( Eliza | herself ) —
d. Grace *verbs* ( Grace | herself ) —
e. Isla *verbs* ( Isla | herself ) —
f. Katherine *verbs* ( Katherine | herself ) —
g. Margaret *verbs* ( Margaret | herself ) —
h. Neha *verbs* ( Neha | herself ) —
i. Patricia *verbs* ( Patricia | herself ) —
j. Rachael *verbs* ( Rachael | herself ) —
k. Tracy *verbs* ( Tracy | herself ) —
l. Ursula *verbs* ( Ursula | herself ) —
m. Winnifred *verbs* ( Winnifred | herself ) —

The training set then included all *Person*-reflexive sentences with masculine antecedents and the lone remaining *Person*-reflexive sentences with a feminine antecedent, "Yvette", as shown below in (37).

(37) Yvette *verbs* ( Yvette | herself ) [present in Alice-$\varepsilon$-14]

**Summary of findings:** Rather than look at any individual experiment in the -$\varepsilon$ series, it will be more instructive to consider how models' performance is affected as the number of (feminine) contexts withheld increases.

One notable observation about the entire family of experiments is the realtive performance of the various model structures. In contrast to the -$\delta$ and -$\gamma$ experiments, where `SRN`s and `GRU`s with attention attained the best performance, the -$\varepsilon$ family shows that `SRN`s, in general, fail at this task even when aided by attention. Conversely, `GRU`s and `LSTM`s prove quite capable of attaining high accuracy even when all but one feminine context is withheld during training.

Of additional interest is the impact that attention has on model performance. The -$\varepsilon$ family shows that sometimes, attention is *not* all you need, and in fact may prove detrimental. While previous experiments showed how attention could improve model performance we see here that attention *negatively* impacts the accuracies of `GRU`s; `GRU`s sans attention outperform those with attention on high-number-withheld training sets.

Finally, we see a case where `LSTM`s have excellent performance when compared to other architectures. Both attentative and non-attentative variants of `LSTM`s have accuracies above 90%, on par with the non-attentative `GRU`s.

This set of experiments shows how qualitatively different `difficult` probelms can be for networks when compared to the -$\gamma$/-$\delta$ cases. The differences between the effects that structure and attention have on the models' success at these tasks is notable since it shows that attaining high accuracy is not necessarily universal even within the same problem domain; making the problem "more difficult" in one way may impact models differently than in another way.

From a theoretical standpoint, its worth thinking about what it meants to generalize to new lexical antecedents in this case. Why is it that the models' performance on Alice-reflecxive sentences is affected at all by the presence or absence of reflexive sentences with other antecedents? Obviously this is the main question we care about in terms of lexical generalization, but undestanding what leads to the empirical findings here is worth considering.

One last thing of note in this family of experiments is an examination of how quickly models learn the generalization task relative to (1) other structures, (2) the validation set, and (3) the individual names in the generalization task. We observe that `GRU`s without attention and both flavors of `LSTM`s learn to solve the generalization task *before* they solve the full validation set, shown by the fact that accuracy on the generalization set increases before accuracy on the validation set. We also note that models with attention learn the generalization task (and the validation task) within a shorter amount of time that those without; this is characteristic of learning to generalize "all at once," rather than learning each separate interpretation of "herself" individually.

## 4.8 Alice-$\zeta$: Who's Alice and who's Claire? part II: electric boogaloo

In Alice-$\zeta$ we repeat the process of Alice-$\epsilon$ but withold combinations of masculine and feminine names instead of only witholding feminine names. This is to test if there is any underlying connection between the reflexives "himself" and "herself". Clearly speakers of English will undestand that these anaphors differ only in $\varphi$-features, but networks have no such underlying knowledge or biases. In principle, these could be totally separate tokens which do not impact each other in any way.

**Alice-ζ-2:** We withhold *Person*-reflexive sentences where "Alice" and "Bob" are the antecedents, shown below in (38).

(38) a. Alice *verbs* ( Alice | herself ) [withheld in Alice-ζ-2]
b. Bob *verbs* ( Bob | himself ) —

**Alice-ζ-4:** We withhold *Person*-reflexive sentences where "Alice", "Bob", "Claire", and "Daniel" are the antecedents, shown below in (39).

(39) a. Alice *verbs* ( Alice | herself ) [withheld in Alice-ζ-4]
b. Bob *verbs* ( Bob | himself ) —
c. Claire *verbs* ( Claire | herself ) —
d. Daniel *verbs* ( Daniel | himself ) —

**Alice-ζ-6:** We withhold *Person*-reflexive sentences where "Alice", "Bob", "Claire", "Daniel", "Eliza", and "Francis" are the antecedents, shown below in (40).

(40) a. Alice *verbs* ( Alice | herself ) [withheld in Alice-ζ-6]
b. Bob *verbs* ( Bob | himself ) —
c. Claire *verbs* ( Claire | herself ) —
d. Daniel *verbs* ( Daniel | himself ) —
e. Eliza *verbs* ( Eliza | herself ) —
f. Francis *verbs* ( Francis | himself ) —

**Alice-ζ-14:** We withhold *Person*-reflexive sentences where "Alice", "Bob", "Claire", "Daniel", "Eliza", "Francis", "Grace", "Henry", "Isla", "John", "Katherine", "Lewis", "Margaret", and "Oscar" are the antecedents, shown below in (41).

(41) a. Alice *verbs* ( Alice | herself ) [withheld in Alice-ε-14]
b. Bob *verbs* ( Bob | himself ) —
c. Claire *verbs* ( Claire | herself ) —
d. Daniel *verbs* ( Daniel | himself ) —
e. Eliza *verbs* ( Eliza | herself ) —
f. Francis *verbs* ( Francis | himself ) —
g. Grace *verbs* ( Grace | herself ) —
h. Henry *verbs* ( Henry | himself ) —
i. Isla *verbs* ( Isla | herself ) —
j. John *verbs* ( John | himself ) —
k. Katherine *verbs* ( Katherine | herself ) —
l. Lewis *verbs* ( Lewis | himself ) —
m. Margaret *verbs* ( Margaret | herself ) —
n. Oscar *verbs* ( Oscar | himself ) —

**Alice-ζ-24:** We include only the *Person*-reflexive sentences where where "Yvette" or "Xerxes" serve as the antecedents, as shown below in (42).

(42) a. Yvette *verbs* ( Yvette | herself ) [present in Alice-$\varepsilon$-24]
b. Xerxes *verbs* ( Xerxes | himself ) —

**Summary of findings:** The -$\zeta$ experiemnts turned out much the same as the -$\varepsilon$ experiments did, though with a "delay". Since each of the $\zeta$-*n* cases had twice as many examples of masculine/feminine reflexive contexts as the $\varepsilon$-*n* cases, the $\zeta$-*n* had better performance than the corresponding $\varepsilon$-*n* runs. In general, witholding masculine names doesn't seem to make much of an impact on accuracy for feminine contexts, and vice-versa.

# 5 References

[1] R. Frank, D. Mathis, and W. Badecker. *The Acquisition of Anaphora by Simple Recurrent Networks*. 18 June 2013. *Language Acquisition*.