Learning Pronoun Case from Distributional Cues: Flexible Frames for Case Acquisition

Xiaomeng Ma[1], Martin Chodorow[1,2], Virginia Valian[1,2]
[1]The Graduate Center and [2]Hunter College, CUNY
xma3@gradcenter.cuny.edu, mchodorow@hunter.cuny.edu, vvalian@gc.cuny.com

Abstract

Case is an abstract grammatical feature that indicates argument relationship in a sentence. In English, cases are expressed on pronouns, as nominative case (e.g. *I*, *he*), accusative case(e.g. *me*, *him*) and genitive case (e.g. *my*, *his*). Children correctly use cased pronouns at a very young age. How do they acquire abstract case in the first place when different cases are not associated with different meanings? This paper proposes that the distributional patterns in parents' input could be used to distinguish grammatical cases in English.

1 Introduction

Case is a special grammatical property of nouns, pronouns, adjectives, participles or numerals whose value reflects the grammatical function performed by that word in a phrase, clause or sentence. In some languages, all of the word categories mentioned above take different inflected forms depending on their case. English, however, has largely lost its inflected cases and only expresses three cases on personal pronouns: nominative case (e.g. *I, he, she*), accusative case (e.g. *me, him*) and genitive case (e.g. *my, his, her*). These cases are used to mark different relationships between arguments and are commonly referred to as abstract case. For example, nominative pronouns are used as the subject of the sentence; accusative pronouns are used as the objects; and genitive pronouns are used as determiners. Case is formally assigned by the +FINITE feature in the syntactic projection.

English-speaking children are able to use cased pronouns correctly at a very early age. However, between ages 2-4, they reportedly make pronoun case errors such as 'where does *him* go'[1], 'all of *they* going go in here'[2] and 'what *my* doing'[3]. These errors might provide insights into how children acquire grammatical case. For over 40 years, researchers have proposed different explanations of abstract case acquisition based on such errors (e.g. Huxley, 1970; Budwig, 1989; Rispoli, 2005; Fitzgerald et al., 2017). The syntactic explanation argues that children's knowledge of grammatical case is the result of syntactic maturation of the tense and agreement system (Vainikka, 1993; Wexler, 1994; Schütze and Wexler, 1996). The morphosyntactic theory proposes that children form a paradigm for each pronoun, including features like case, person, gender and number and retrieve different forms for different contexts (Rispoli, 1994, 1998).

In addition, parents' input has been implicated to play an important role in case acquisition. Previous studies have argued that some of the children's pronoun case errors could be explained by the ambiguous uses of pronouns in parents' input. For example, Pelham (2011) argued that English-speaking children made more pronoun case errors than German-speaking children because there are more case-ambiguous pronouns (e.g. *you* and *it*) in English. Tomasello (2000) suggested that children could be confused by phrases such as 'Let *me* do it', thus producing errors like '*me* do it'.

Are children able to learn pronoun case in the face of ambiguity? We investigated whether parents' input is informative enough for children to learn pronominal case. In English, nominative, accusative, and genitive case have different distributional patterns. For example, nominative pronouns are more likely to occur before a verb (e.g. '*I* see'), whereas accusative or genitive pronouns

[1]Utterance from Becky at 2;6 in the Manchester corpus (Theakston et al., 2001), Manchester/Becky/020619.cha

[2]Utterance from Nina at 2;11 in the Suppes corpus (Suppes, 1974), Suppes/021021.cha

[3]Utterance from Eve at 2;1 in the Brown corpus (Brown, 1973), Brown/Eve/020100b.cha

rarely appear before a verb (e.g. '**me/my* see'). However, cases do not have exclusive distributional patterns. Some patterns are shared by more than one case: an accusative pronoun case precede a verb in phrases like 'let *him* go' or 'help *me* see.' (Tomasello, 2000). Therefore, we asked if patterns of word co-occurrence could be used to differentiate pronoun cases. In the section 2, we review prior approaches using co-occurrence patterns for word categorization. In section 3, we introduce our methods and models. In section 4, we explain the setup and results of three analyses. Section 5 summarizes our contributions.

2 Related Work

Distributional information can be effective in grammatical categorization tasks. Redington et al. (1998) demonstrated that the context of a target word, including the previous words and following words, can be used to cluster the target word into different grammatical categories. Their model achieved high accuracy in categorization in general; however, for words that appear in less frequent contexts, accuracy suffered. Mintz (2003) proposed that frequent local trigram frames consisting of one word before the target word and one word following it (an `aXb` frame, where `X` is the target word) contain enough information for grammatical categorization. For example, in the frame 'to `X` to', `X` is likely to be a verb, e.g. 'to go to'. Mintz examined the 45 most frequent `aXb` frames in parents' input and showed that the accuracy for `X`'s grammatical categorization was over 0.90. However, only a small portion of words appear in the frequent `aXb` frames. In order to categorize more words, Clair et al. (2010) separated the `aXb` frame into two bigram frames: `aX + Xb`. They suggested that instead of learning the co-occurring frame 'a_b' as a whole unit, it is more efficient to treat it as two flexible bigrams 'a_' and '_b' which are more useful in learning. They trained feedforward neural networks on 100,000 samples of the `aXb` frame and the `aX + Xb` frame had better categorization accuracy (0.73) than the `aXb` frame (0.53).

Grammatical cases are similar to grammatical categories in that both reflect certain syntactic features of the word. In this paper, we trained models to predict the pronoun case of `X` using `aXb` and `aX + Xb` frames. The purpose of the study is not to provide a model to explain children's grammatical case acquisition, but to examine if the distributional patterns in parents' input are informative enough to distinguish pronoun cases. The results do not indicate whether children acquire pronoun case from parents' input but suggest a possible source from which children could learn pronoun case.

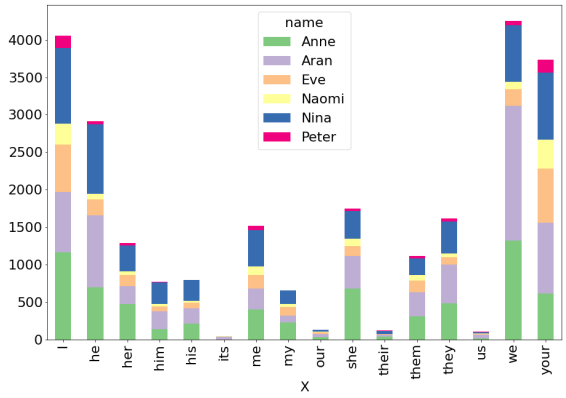

Figure 1: Pronoun tokens by the parents of six children

3 Methods

3.1 Corpus Summary

Following Mintz (2003) and Clair et al. (2010), we used the same six corpora of child-directed speech from CHILDES (MacWhinney, 2014): Anne and Aran (Theakston et al., 2001), Eve (Brown, 1973), Naomi (Sachs, 1983), Nina (Suppes, 1974), Peter (Bloom et al., 1974). We analyzed the utterances in the files where the child is younger than 2;6 years old. The pronouns were extracted with the part-of-speech tags assigned by the MOR parser (MacWhinney, 2012) in CHILDES: `pro:sub` for nominative pronouns, `pro:obj` for accusative pronouns and `det:poss` for genitive pronouns. Case-ambiguous pronouns 'you' and 'it' were excluded from the study since they were tagged as `pro:per` in all argument positions. The pronoun 'her' was included since it was tagged as `pro:obj` for its accusative use and `det:poss` for its genitive use. Each pronoun was extracted with its `aXb` context. Table 1 summarizes the number of tokens of all pronouns and each case, and the number of types for `aX`, `Xb` and `aXb`. Figure 1 shows the token frequencies of the pronouns produced by the children's parents.

3.2 Model Architecture

We used supervised learning with a feedforward connectionist model to compare the accuracy of the `aXb` model and the `aX + Xb` model. For the `aXb`

	Nominative	Accusative	Genitive	Pronoun Tokens	aX types	Xb types	aXb types
Aran	4518	1014	1454	6986	445	927	2489
Anne	4343	1080	1392	6815	428	707	2308
Eve	1292	479	1029	2800	278	500	1364
Naomi	599	249	503	1352	224	364	806
Nina	3490	1195	1571	6256	400	747	2376
Peter	339	135	207	681	187	250	475
Total	14581	4152	6156	24889	898	1672	7355

Table 1: Token counts of three pronoun cases and type counts of three context frames

model, the input consisted of one one-hot vector, representing '`a_b`'. For the `aX + Xb` model, the input consisted of two one-hot vectors, representing '`a_`' and '`_b`' respectively. The two models are shown in Figure 2 and Figure 3, with 'let me do' as an example. For the `aXb` model, the input unit represents 'let_do'. For the `aX + Xb` model, one input represents 'let_' and the other input represents '_do'. The connectionist model used the following parameters: (1) number of hidden units was set to 200 and initialized randomly for each model; (2) the non-linearity was relu.

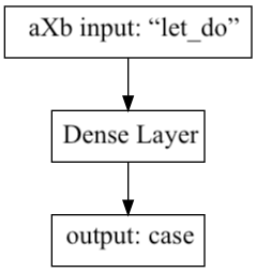

Figure 2: The architecture of `aXb` model

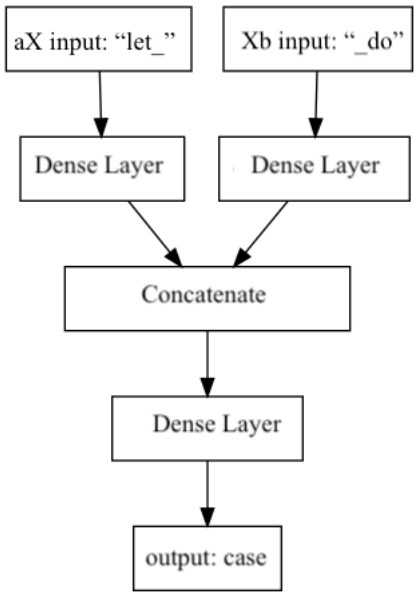

Figure 3: The architecture of `aX + Xb` model

3.3 Evaluation

We use the classification accuracy for each case to compare the `aXb` model and `aX + Xb` models. In addition, following Clair et al. (2010) we also report the asymmetric lambda value Goodman and Kruskal (1979) to evaluate the association among the classification of grammatical cases. Lambda is defined as the proportional reduction in prediction error. It provides insight into the extent to which the model's prediction is based on the actual category. Lambda is in the range of [0,1], where 0 indicates there is no association between predicted and actual categories, and 1 indicates a perfect association. For example, if the model categorizes all frames as nominative cases simply because that is the most frequent case, then the accuracy will be 0.586 (14581/24889), but the lambda will be 0.

3.4 Training and Testing

We measured and compared the classification accuracy of models by applying 10-fold cross validation on the union of the six children's corpora. The `aXb` model and `aX + Xb` model were trained using the same 10-fold cross-validation split. All the frames were used for both training and testing.

We used the Adam optimization algorithm to minimize the mean squared error (MSE) loss function over the training data. We trained the model on a maximum of 100 epochs with a batch size of 32. Early stopping methods were applied to stop the training when the accuracy did not change in 10 consecutive training rounds.

4 Experiments

4.1 Experiment 1: Models `aXb` vs `aX + Xb` in Categorizing Grammatical Cases

Method. All models were trained and evaluated following the steps in Section 3. Following (Clair et al., 2010), we split the training of each model into a token-training phase and a type-training

phase. In the token-training phase, we trained the models on all 24889 pronoun patterns. In the type-training phase, we trained the models only on 7355 tokens of unique aXb types.

Results of two training phases Tables 2 and 3 show the overall classification accuracies and lambda scores of aXb and aX + Xb on each child's corpus. Both models achieved very high accuracy with 24889 tokens. In addition, the lambda scores showed that almost perfect associations, suggesting that the aXb and aX + Xb models are very effective in predicting the correct grammatical case. Figures 4 and 5 show the heatmaps of the classification results. All three cases are classified with high accuracy. The heatmaps also indicate that the two models make different classification errors. For example, for genitive case, the aX + Xb model is more likely to miscategorize it as an accusative case whereas the aXb model is more likely to label it as a nominative case.

	aX + Xb		**aXb**	
	Accuracy	λ	**Accuracy**	λ
Aran	0.984	0.956	0.962	0.894
Anne	0.984	0.957	0.962	0.897
Eve	0.979	0.961	0.960	0.928
Naomi	0.983	0.969	0.951	0.914
Nina	0.987	0.970	0.951	0.911
Peter	0.982	0.965	0.954	0.913
Total	**0.984**	**0.962**	**0.960**	**0.907**

Table 2: Results of training on 24889 total tokens

	aX + Xb		**aXb**	
	Accuracy	λ	**Accuracy**	λ
Aran	0.968	0.940	0.849	0.631
Anne	0.963	0.936	0.841	0.639
Eve	0.968	0.931	0.872	0.648
Naomi	0.953	0.902	0.878	0.708
Nina	0.974	0.952	0.834	0.600
Peter	0.963	0.927	0.827	0.619
Total	**0.967**	**0.939**	**0.847**	**0.631**

Table 3: Results of Training on 7355 tokens of unique types

When the sample size drops to 7355 tokens, the accuracies and the lambda scores also change for both models. The performance of aX + Xb is not heavily affected by a smaller sample size: the accuracy changes from 0.984 to 0.967 and the lambda score changes from 0.962 to 0.939. In contrast,

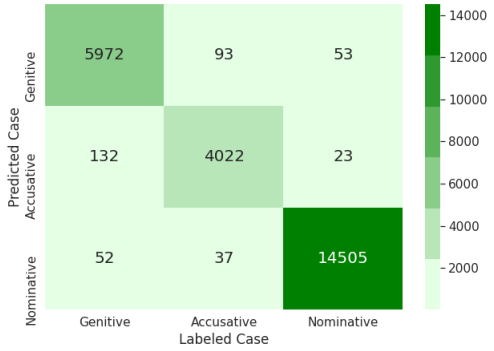

Figure 4: Heatmap of each case's classification in aX + Xb model of 24889 total tokens

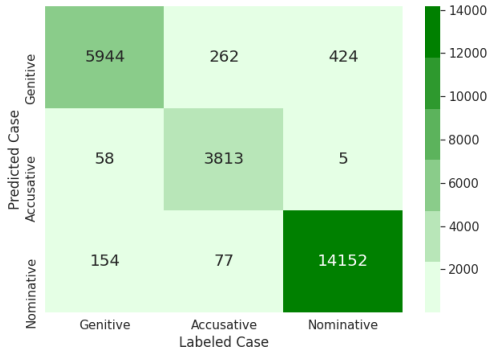

Figure 5: Heatmap of each case's classification in aXb model of 24889 total tokens

the aXb model shows a large decline in the accuracy and the lambda score when the sample size drops: the accuracy falls to 0.847 and the lambda score drops to 0.631. Thus, aX + Xb not only has higher accuracy, but also is less vulnerable to small sample size. Figures 6 and 7 are the classification heatmaps of each case. We also plotted the classification accuracies of each pronoun for each child's input, which can be found in Figures 11 - 14 in the Appendix.

4.2 Experiment 2: Predicting the Pronoun Using aX + Xb Model with Person, Gender, Number Information

Method. Since the aX + Xb model achieved high accuracy in grammatical case classification, we asked if the pronoun can be effectively predicted when person, gender and number information are given in training. In the second experiment, we coded the person, gender, and number information for each pronoun (e.g. 'he' would be coded as third-person, masculine and singular) and added it as an additional input to the aX + Xb model. We

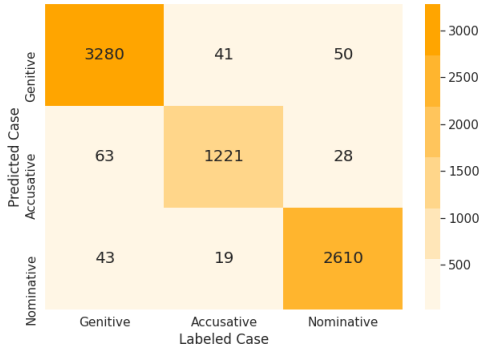

Figure 6: Heatmap of each case's classification in `aX + Xb` model of 7355 tokens of unique types

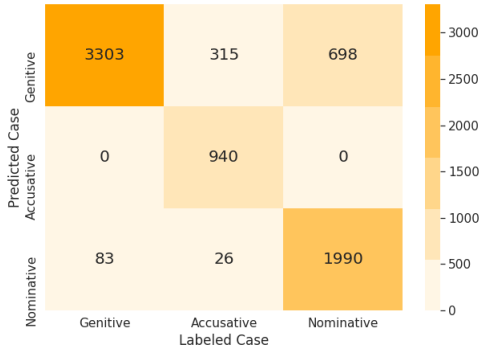

Figure 7: Heatmap of each case's classification in `aXb` model of 7355 tokens of unique types

trained the model on the same 24889 total tokens and 7355 tokens of unique types as in Experiment 1, and used the same 10-fold cross-validation splits.

Results. Table 4 shows the results. With gender, person and number as additional input, the `aX + Xb` model can predict the pronoun at an accuracy of 0.994 with 24889 total tokens and 0.982 with 7355 tokens of unique types. In addition, the lambda scores indicate an almost perfect association. Given additional information on person, gender, and number, the `aX + Xb` model is extremely effective in predicting the pronoun.

Figures 8 and 9 show the classification results for each pronoun. The classification errors on each pronoun are usually case errors (e.g. 'I' mislabeled as 'me' or 'my'). There are few errors on the gender (e.g. 'he' mislabeled as 'she') and almost no errors on number and person. Each child's pronoun accuracy in shown in Figures 15 and 16 in the Appendix.

	24889 tokens		**7355 types**	
	Accuracy	λ	**Accuracy**	λ
Aran	0.994	0.992	0.980	0.971
Anne	0.994	0.992	0.980	0.976
Eve	0.993	0.990	0.983	0.972
Naomi	0.993	0.995	0.980	0.967
Nina	0.996	0.994	0.987	0.982
Peter	1.000	1.000	0.983	0.975
Total	**0.994**	**0.993**	**0.982**	**0.975**

Table 4: Results of `aX + Xb` model Predicting Pronoun on 24889 and 7355 tokens with gender, number, person information

4.3 Experiment 3: Corpus Analysis of Children's Pronoun Case Error Patterns

Experiments 1 and 2 have show that distributional patterns are extremely effective in pronoun case categorization, suggesting that parents' input is informative for pronoun case learning. In experiment 3, we examine how well children learn pronouns. We conducted a corpus analysis of all six children's utterances and calculated their pronoun case errors.

Methods. We searched the pronoun case errors in each child's utterances in all available files (including those with an age older than 2;6) in the corpora. To identify pronoun case errors, the part-of-speech tags in CHILDES were used. For English data, the automated annotation system has been reported to have high-level accuracy: the MOR program reaches 97% accuracy in word categorization, and the GRASP program has 95.8% accuracy in determining the subject in the sentence and 94.1% accuracy in determining the object in the sentence (MacWhinney, 2012; Sagae et al., 2010). After the errors were first located using the MOR program and GRASP programs, two annotators independently hand-checked the errors.

Results. Table 5 shows the accuracy for each child's pronoun case use. Most children have very high accuracy in their pronoun case uses, except for Nina, whose accuracy is 0.926. The overall pronoun case accuracy for all 6 children is 0.97, which is similar to the results of the `aX + Xb` model (0.967 for unique types and 0.984 for total tokens).

Figure 10 shows children's errors on pronoun cases. Children's errors are different from the classification models' errors. Children never mistreated a genitive pronoun or a nominative pronoun as an accusative pronoun.

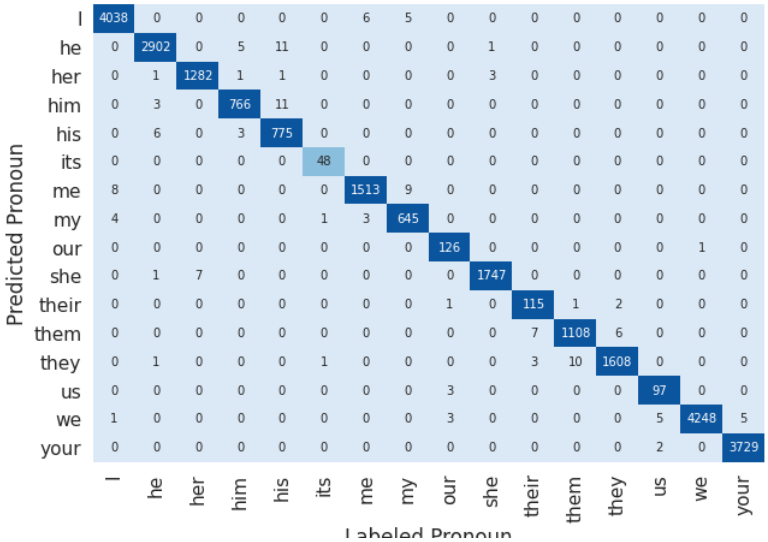

Figure 8: Heatmap of pronoun classification results on 24889 total tokens

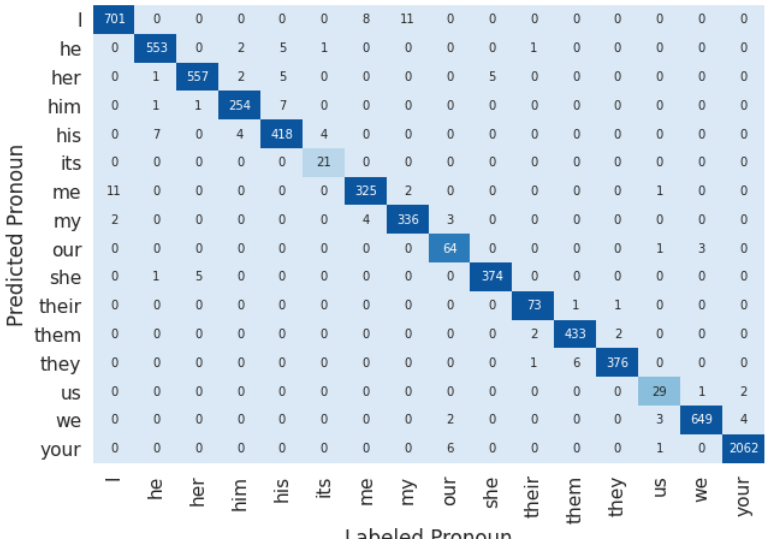

Figure 9: Heatmap of pronoun classification results on 7355 tokens of unique types

5 Discussion and Conclusion

In this work, we proposed that distribution patterns could be used to distinguish pronoun cases. We trained models on the fixed trigram frame `aXb` and flexible frame `aX + Xb` with a large sample size and a smaller one. The results showed that the distributional patterns are extremely effective in categorizing grammatical case of pronouns. Based on the high accuracy results with case categorization, we further explored pronoun categorization with person, gender, and number information as additional input. With large sample size, our model achieved almost perfect pronoun categorization accuracy. We then conducted a corpus analysis to examine children's pronoun case acquisition. Most of the children have a similar accuracy rate as our training model.

Our experiments showed that distributional patterns in parents' input are very useful in categorizing grammatical cases. Our model showed a similar accuracy rate as children's real-life pronoun case acquisition. However, the similar accuracy rate does not demonstrate that children actually utilize distributional patterns in learning and the differences between the errors made by training models and by children suggest children may be

	Errors	Total Pronouns	Accuracy
Anne	57	5009	0.989
Aran	25	8450	0.997
Peter	115	4077	0.971
Eve	49	2685	0.982
Naomi	64	3249	0.980
Nina	633	8609	0.926
Total	**943**	**32079**	**0.970**

Table 5: Results of each child's pronoun case errors and accuracy

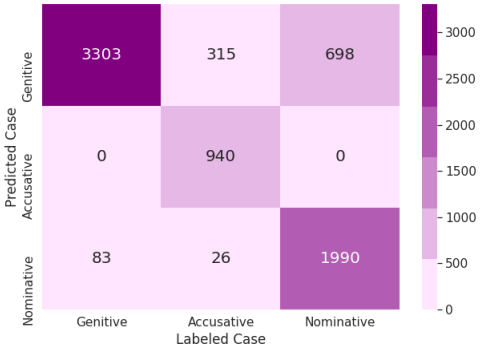

Figure 10: Heatmap of pronoun case uses by children

using a different procedure or an additional procedure. Further investigations of the classification errors and children's pronoun case errors will be informative for understanding the process of case categorization.

References

Lois Bloom, Lois Hood, and Patsy Lightbown. 1974. Imitation in language development: If, when, and why. *Cognitive psychology*, 6(3):380–420.

Roger Brown. 1973. *A first language: The early stages.* Harvard U. Press.

Nancy Budwig. 1989. The linguistic marking of agentivity and control in child language. *Journal of Child Language*, 16(2):263–284.

Michelle C St Clair, Padraic Monaghan, and Morten H Christiansen. 2010. Learning grammatical categories from distributional cues: Flexible frames for language acquisition. *Cognition*, 116(3):341–360.

Colleen E Fitzgerald, Matthew Rispoli, and Pamela A Hadley. 2017. Case marking uniformity in developmental pronoun errors. *First Language*, 37(4):391–409.

Leo A Goodman and William H Kruskal. 1979. Measures of association for cross classifications. In *Measures of association for cross classifications*, pages 2–34. Springer.

Renira Huxley. 1970. The development of the correct use of subject personal pronouns in two children. In Flores D'arcais, G.B., and Levelt, W. J. M. (eds.). *Advances in Psycholinguistics*, pages 141–165.

Brian MacWhinney. 2012. Morphosyntactic analysis of the childes and talkbank corpora. In *LREC*, pages 2375–2380.

Brian MacWhinney. 2014. *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press.

Toben H Mintz. 2003. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1):91–117.

Sabra D Pelham. 2011. The input ambiguity hypothesis and case blindness: an account of cross-linguistic and intra-linguistic differences in case errors. *Journal of Child Language*, 38(2):235–272.

Martin Redington, Nick Chater, and Steven Finch. 1998. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22(4):425–469.

Matthew Rispoli. 1994. Pronoun case overextensions and paradigm building. *Journal of Child Language*, 21(1):157–172.

Matthew Rispoli. 1998. Me or my: Two different patterns of pronoun case errors. *Journal of Speech, Language, and Hearing Research*, 41(2):385–393.

Matthew Rispoli. 2005. When children reach beyond their grasp: Why some children make pronoun case errors and others don't. *Journal of Child Language*, 32(1):93–116.

Jacqueline Sachs. 1983. Talking about the there and then: The emergence of displaced reference in parent-child discourse. *Children's Language*, 4:1–28.

Kenji Sagae, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner. 2010. Morphosyntactic annotation of childes transcripts. *Journal of Child Language*, 37(3):705–729.

Carson Schütze and Kenneth Wexler. 1996. Subject case licensing and English root infinitives. In *Proceedings of the 20th annual Boston University conference on language development*, volume 2, pages 670–681. Cascadilla Press Somerville, MA.

Patrick Suppes. 1974. The semantics of children's language. *American Psychologist*, 29(2):103.

Anna L Theakston, Elena VM Lieven, Julian M Pine, and Caroline F Rowland. 2001. The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, 28(1):127–152.

Michael Tomasello. 2000. Do young children have adult syntactic competence? *Cognition*, 74(3):209–253.

Anne Vainikka. 1993. Case in the development of english syntax. *Language Acquisition*, 3(3):257–325.

Ken Wexler. 1994. 14 optional infinitives, head movement and the economy of derivations1. *Verb Movement*, page 305.

A Appendix

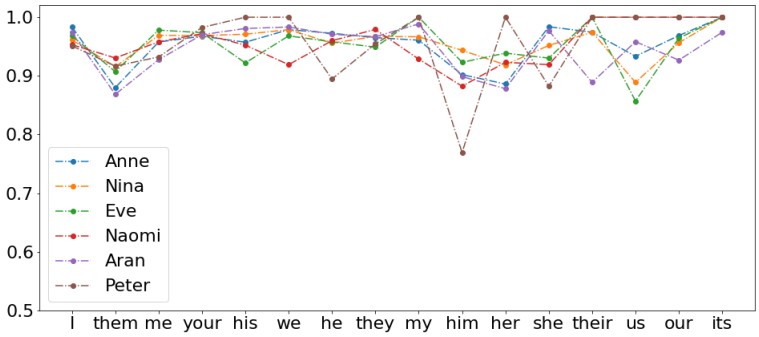

Figure 11: `aXb` model accuracy with 24889 total tokens

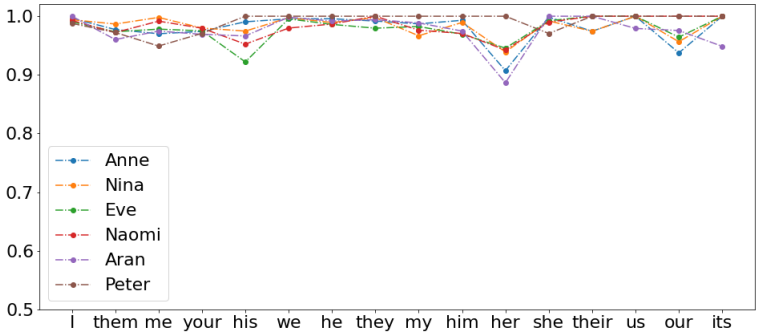

Figure 12: `aX+Xb` model accuracy with 24889 total tokens

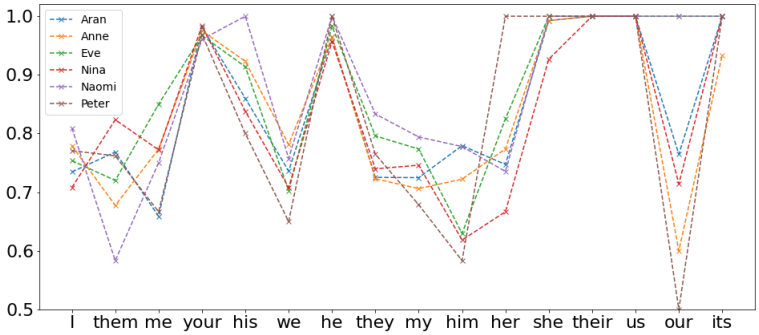

Figure 13: `aXb` model accuracy with 7355 tokens of unique types

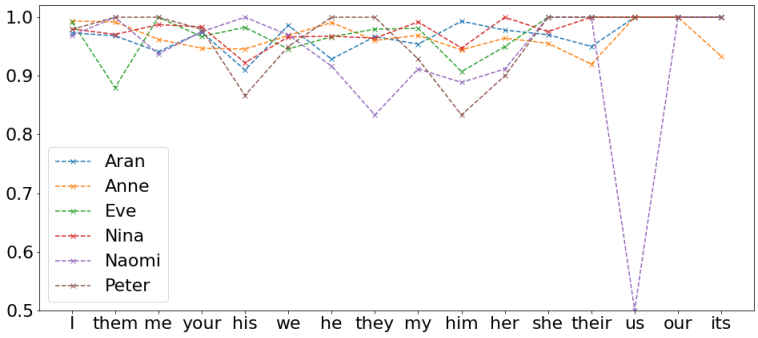

Figure 14: `aX + Xb` model accuracy for each pronoun with 7355 tokens of unique types

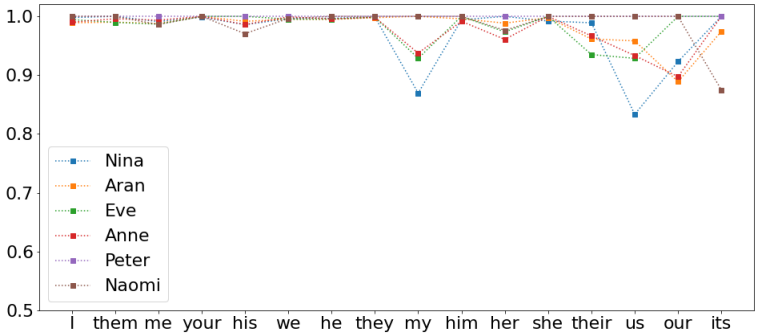

Figure 15: Accuracies of `aX + Xb` model with person, gender, number information for each pronoun with 24889 total tokens

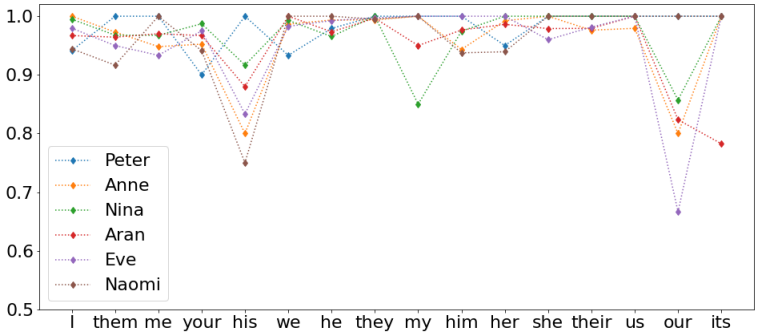

Figure 16: Accuracies of `aX + Xb` model with person, gender, number information for each pronoun with 7355 tokens of unique types