

000
001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047

The emergence of number and syntax units in LSTM language models

048
049
050
051
052
053
054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095

Anonymous NAACL submission

Abstract

Recent work has shown that LSTMs trained on a generic language modeling objective capture syntax-sensitive generalizations such as long-distance number agreement. We have however no mechanistic understanding of how they accomplish this remarkable feat. Some have conjectured it depends on heuristics that do not truly take hierarchical structure into account. We present here a detailed study of the inner mechanics of number tracking in LSTMs at the single neuron level. We discover that number information is managed by very few “grandmother cells” in a localist fashion. Importantly, the behaviour of the number cells is partially controlled by other units that are independently shown to track the syntactic structure of sentences. We conclude that LSTMs are, to some extent, implementing genuinely syntactic processing mechanisms, paving the way to a more general understanding of grammatical encoding in LSTMs.

1 Introduction

[1-6]

Studies showing that LSTMs trained on language modeling do well on the agreement task: (?; ?), to a lesser extent: (?; ?). Studies conjecturing this is just heuristics: (?; ?). Grandma cells: (?).

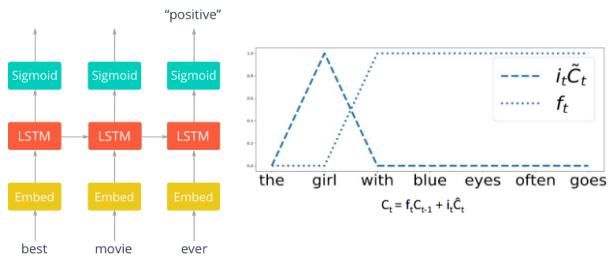


Figure 1: Caption.

2 Related literature

2.1 Interpreting LSTM networks

Short survey of works that:

- Interpret neural networks in general (e.g., CNNs in vision)
- Interpret LSTMs (e.g. Karpathy’s)
- Interpret LSTM language models

2.2 Subject-verb agreement in English

- Subject-verb agreement from psycholinguistics (e.g., Miller and Bock, Franck and Rizzi) - SV agreement in LSTMs (Linzen 2016, Gulordava 2018, etc.). - Relate to Nelson et. al 2017 PNAS, an intracranial study that identifies electrodes whose high-gamma activity correlates with syntactic tree-depth (number of open nodes)

3 Data

We generated number-agreement tasks (NATasks) with fixed syntactic structures and varied lexical material that probe subject-verb number agreement in increasingly challenging conditions **setups**. The different structures are illustrated in Table 1 by examples where all forms are in the

096	Simple	the boy greets the guy
097	Adv	the boy probably greets the guy
098	2Adv	the boy most probably greets the guy
099	CoAdv	the boy openly and deliberately greets the guy
100	NamePP	the boy near Pat greets the guy
101	NounPP	the boy near the car greets the guy
102	SubjRel	the boy that avoids the girl greets the guy
103	ObjRel	the boy that the girl avoids greets the guy
104	ObjRel0	the boy the girl avoids greets the guy

Table 1: Number agreement data-sets illustrated by representative singular sentences.

singular. Distinct sentences were randomly generated by selecting (different) words from pools of 20 subject/object nouns, 15 verbs, 10 adverbs, 5 prepositions, 10 proper nouns and 10 location nouns. The items were selected so that their combination would not lead to semantic anomalies. For each NA-task, we generated singular and plural versions of each sentence. We refer to each such version as a *condition*. For NA-tasks that have other nouns occurring between subject and main verb, we also systematically varied their number, resulting in four conditions instead of two. For example, the NounPP sentence in the table illustrates the SS (singular-singular) condition. The corresponding sentences in the other conditions are: "the boy near the cars greets the guy" (SP), "the boys near the car greet the guy" (PS), and "the boys near the cars greet the guy" (PP).

~~Yair: not sure if this is completely accurate - it's not that the pair was presented, but only the prefix of the sentence (the part upto to the verb), and then the likelihood was compared. I would anyway move this part to the 'model evaluation' section under 'The models' or to the results section: In all cases, ungrammatical versions with (main-clause) subject-verb number mismatches were also generated. For the NounPP example in the table, the mismatched counterpart is "the boy near the car greet the guy". Each correct/mismatched pair was presented to the pre-trained network, and we computed its accuracy by counting as hits all cases where it assigned a higher likelihood to the correct version. Suggestion (Yair): For each NA-task, model performance was evaluated for each of its~~

conditions separately, by computing its accuracy in predicting the correct form of the main verb. Specifically, for each sentence from a condition, the model was presented with all words upto one word before the main verb. If at this point the model assigned higher likelihood to the correct form of the verb then this trial was marked as a hit. Accuracy was then computed by counting the fraction of hits in the entire set of sentences in the condition.

~~Yair: consider move these explanations about contrasts to the results section: Note that 2Adv features the same subject-verb distance as NamePP, but without gender-carrying words acting as possible distractors in the middle. Similarly, CoAdv can serve as a distracto-free control for NounPP.~~

Data-set sizes ranged from 600 (Simple) to 18k sentences (relatives) ~~?Yair: for all conditions taken together or should this number multiplied by the number of conditions?~~, based on the possibilities for variation allowed by the combinatorics given each structure.¹

Finally, we also used the naturalistic, corpus-derived agreement test set of ~~?~~, in the version made available by ~~?~~.

4 The models

I would describe in this section the regression and language models.

4.1 LSTM language model

4.1.1 Architecture and dynamics

~~TODOS: - Number and type (embedding, LSTM, softmax) of layers, including dimensions. - Describe the dynamics of the network (list of equations of the LSTM)~~

4.1.2 Model training and evaluation

~~- Task (refer to Gulordava et. al) - Describe and give (nick) names for the other models, e.g., LM-high-dropout, LM-SEED1 - The model was evaluated on a left-out test set...perplexity..~~

¹We uploaded the data-set generation script as supplementary material.

192	4.2 Regression model	of a verb immediately following a subject. Finally, subsection 5.1.4 characterizes the structure of the efferent weights of number units, which propagate number information to the output layer.	240
193	- Explain that the model was used to predict the depth of the syntactic tree from network activity		241
194			242
195			243
196			244
197	4.2.1 Model description		245
198	- Describe the Features: network activity (hidden/cell activity) - Describe the label: Tree depth (refer to the section describing the synthetic data) - Describe the model: A Ridge/LASSO model.		246
199			247
200			248
201			249
202			250
203	4.2.2 Model training and evalution		251
204	- Model training: nested 5-fold CV procedure.		252
205	Train/val/test. Optimal regularization size was estimated from the validation set (report optimal lamda - figures in FAIRNS.pdf on slack)		253
206	- model evaluation: R-squared on test set. Report resulting values (text+figures in FAIRNS.pdf).		254
207			255
208			256
209			257
210			258
211	5 Results		259
212			260
213	[1]		261
214			262
215	5.1 Long-range number-units		263
216	To successfully perform the NA-task, the LSTM		264
217	network should encode and store the grammatical		265
218	number of the subject up to one step before		266
219	the verb, when prediction of the verb form		267
220	(singular or plural) occurs. In some cases, this		268
221	may be quite challenging, in particular in the		269
222	case of a long-range dependency between subject		270
223	and verb, and when another noun with an oppo-		271
224	site number appears before the verb (cite). This		272
225	section explores the underlying mechanism that		273
226	enables the network to encode and store num-		274
227	ber information in various syntactic construc-		275
228	tions, including ones with an interfering noun,		276
229	and has the following structure: subsection 5.1.1		277
230	describes an ablation study, which reveals <i>long-</i>		278
231	<i>range number units (LR-number units)</i> that can		279
232	store and carry number information from sub-		280
233	ject to verb, also across interfering nouns. Sub-		281
234	section 5.1.2 describes the intricate gate and		282
235	state dynamics of LR-number units during the		283
236	processing of sentences with long-range depen-		284
237	dencies. Section 5.1.3 describes other number		285
238	units that encode grammatical number for only		286
239	short-range dependencies, such as in the case		287

	NA task	C	770	776	988	1283	Full
288	Simple	S	-	-	-	-	100
289	Adv	S	-	-	-	-	100
290	2Adv	S	-	-	-	-	99.8
291	Co-Adv	S	-	-	84.0	84.0	98.8
292	namePP	S	-	-	-	-	98.9
293	nounPP	SS	-	-	-	-	97.5
294	nounPP	SP	-	-	58.8	-	88.5
295	subjrel	SS	-	-	88.0	-	97.0
296	subjrel	SP	-	-	-	-	58.8
297	objrel	SS	-	-	-	-	64.7
298	objrel	SP	-	-	-	-	45.7
299							
300	Simple	P	-	-	-	-	100
301	Adv	P	-	-	-	-	99.6
302	2Adv	P	-	-	-	-	99
303	Co-Adv	P	-	78.9	-	-	99.7
304	namePP	P	-	57.6	-	-	66.8
305	nounPP	PS	85.2	49.7	-	-	93.2
306	nounPP	PP	-	81.7	-	-	98.3
307	subjrel	PS	85.8	58.6	-	-	87.8
308	subjrel	PP	-	88.1	-	-	99.3
309	objrel	PS	-	-	-	-	69.0
310	objrel	PP	-	-	-	-	81.0
311	Linzen	-	?	?	?	?	?

Table 2: Ablation experiments results: Percentage of correct subject-verb agreements in all NA-tasks (section 3.1). Full - non-ablated model, C - condition, S - singular, P - plural. For task with two nouns, SS - singular-singular, SP - singular-plural, PS - plural-singular, PP - plural-plural. Red: singular number units, Blue: Plural number units.

viously reported results in humans and LSTM-LMs (cite). Second, having an interfering noun before the verb, with an opposite number than that of the subject, is clearly a more challenging task for the network - we find that for the nounPP, subjrel and objrel tasks: $ACC_{SP} > ACC_{SS}$ and $ACC_{PS} > ACC_{PP}$. We return to this point in section 5.4. Finally, for long-range dependencies, reliably encoding that the subject is singular is in most cases more difficult than plural. For example, in all the above tasks: $ACC_{SS} < ACC_{PP}$ and $ACC_{SP} < ACC_{PS}$. Interestingly, this singular-plural asymmetry has been reported also in humans (cite). We elaborate on this point in the discussion section.

We next describe several important aspects of the ablation-experiment results. First, in all NA-tasks, only four units from the entire network (1300 LSTM units in total) had a significant effect on task performance. This result suggests a local coding scheme for long-range grammatical-number information (TODO: quantify a 'significant' reduction, and perhaps link to the emergence of local coding in neural-network simulations (Bowers) and to findings about grandmother neurons in humans (e.g., Fried)). Second, we note that all number units emerged at the second layer of the network. This seems appropriate if number information needs to be directly projected to the output layer for correct verb-form prediction. In section 5.1.4 we further explore the projection weights from number units. Third, for simple, 1Adv and 2Adv NA-tasks, none of the units had a significant effect on task performance. This suggests that for short-range dependencies number information may be also encoded elsewhere in the network, perhaps via a more distributed code. We therefore make a distinction between long-range (LR) and short-range (SR) number units in what follows. We return to this point in section 5.1.3 (TODO: complete the identification of short-range number units from the resulting weights of the classifier in the generalization-across-time experiment). Fourth, LR-number units can be further divided into two types, depending on the grammatical number of the subject. Units 770 and 776 had a significant effect only when the first noun was plural, but not singular, and vice versa for units 988 and 1283 (blue and red in table 1, respectively). We therefore refer to the former as *plural units* and to the latter as *singular units*. Finally, we note that two of the number units (776 & 988) had an exceptional effect on network performance in both nounPP-SP&PS conditions. These two conditions are in particular revealing since they involve both a long-range dependency (over a prepositional phrase) and an interfering noun before the verb, while performance of the non-ablated network is still relatively high (88.5%&93.2%, respectively) in contrast to these conditions in subjrel and objrel. Ablating one of these two units

384 brought the network from high performance on
 385 the NA-task to around chance-level performance
 386 (58.8% & 49.7%, respectively). In the next section,
 387 we therefore focus on these two units when
 388 exemplifying gate and state dynamics of number
 389 units.
 390

391 5.1.2 Visualizing gate and cell-state 392 dynamics

393 Results from the ablation study suggest that
 394 there's a small set of units that encodes number
 395 information for long-range dependencies, in par-
 396 ticular, we find that in some conditions two units
 397 can bring the network from relatively high per-
 398 formance to around chance-level performance on
 399 the NA-task (section 5.1.1). However, it remains
 400 unclear what is the exact mechanism underly-
 401 ing successful trials in the NA-task, and what
 402 goes wrong in unsuccessful ones. To better un-
 403 derstand this, we now look into gate and state
 404 dynamics of these units during the processing of
 405 sentences from the nounPP NA-task.

406 To anticipate the results and facilitate their
 407 interpretations, we begin by discussing what
 408 could be a solution to the NA-task implemented
 409 by number units in their gate and state activ-
 410 ity. We recall that the update rule of the LSTM
 411 cell has two terms (equation 1.x). In the first
 412 term $f_t * C_{t-1}$, the forget gate controls whether
 413 to keep the previous content C_{t-1} stored in the
 414 cell ($f_t = 1$ - perfect remembering), or forget it
 415 ($f_t = 0$ - complete forgetting). In the second
 416 term $i_t * \tilde{C}_t$, the input gate controls whether the
 417 information currently presented to the network
 418 could be updated onto the cell state: $i_t = 1$ -
 419 full access, $i_t = 0$ - no access. Therefore, to pro-
 420 duce correct number agreement, it seems that
 421 number units should at least have the follow-
 422 ing three properties: (1) The grammatical num-
 423 ber of the subject should first be encoded by
 424 $\tilde{C}_{t_{subject}}$, encoding singular and plural with *dif-*
 425 *ferent* values. (2) To grant the encoded gram-
 426 matical number $\tilde{C}_{t_{subject}}$ access to the cell, the
 427 input gate should be open at the time when the
 428 subject is presented: $i_{t_{subject}} > 0$, and ideally
 429 $i_{t_{subject}} = 1$. In addition, to protect the stored
 430 grammatical number from interfering informa-
 431 tion updating onto the cell, such as in the case

432 of an interfering noun, the input gate should be
 433 closed during all successive time steps until the
 434 verb: $i_t = 0, t < t_{verb}$; (3) Finally, to successfully
 435 store number information in the cell for a long-
 436 range dependency, the forget gate should be in
 437 a remembering state, starting one time step af-
 438 ter the subject: $f_t = 1, t > t_{subject}$. In addition,
 439 to clean up the cell from previously stored
 440 information, the forget gate should reset when
 441 the subject is presented: $f_{t_{subject}} = 0$. Figure
 442 1B summarizes these three presumably desired
 443 properties.

444 Figure 2 presents the actual gate and state
 445 dynamics of units 776 and 988 during the pro-
 446 cessing of sentences from the nounPP NA-task.
 447 For each unit, we draw the dynamics of the sug-
 448 gestion \tilde{C}_t (panels A-B), input-gate (panels C-
 449 D), forget-gate (panels E-F) and the cell variable
 450 (G-H). For each of these cases, the four condition
 451 (SS, SP, PS and PP) are described in separate
 452 curves. Error-bars represent standard deviation
 453 across 1000 sentences in each condition.

454 We describe the results along the order of the
 455 properties discussed above. First, the values of
 456 the cell suggestions \tilde{C}_t of both units seem to
 457 obey the first property. For the singular unit
 458 988, we find that singular nouns are encoded
 459 with negative values $\tilde{C}_{t_{subject}} = -1$, and plurals
 460 with positive $\tilde{C}_{t_{subject}} = 1$ (panel A), and sim-
 461 similarly for unit 776 (panel B). This shows that
 462 singular and plural nouns are indeed encoded
 463 differently by these units, in accordance with the
 464 results of the ablation study that suggested the
 465 labeling of units 988 and 776 as singular and
 466 plural units, respectively.

467 Second, input-gate dynamics of both number
 468 units seem to correspond to the second prop-
 469 erty described above. Input-gate activity spikes
 470 around the subject and stays approximately zero
 471 for subsequent time steps until the verb. One
 472 difference with respect to the desired property
 473 is the non-zero activity of the input gate at
 474 the time step immediately following the subject.
 475 This may be due to various reasons and requires
 476 further research. One possible explanation for
 477 this is that the network has developed this be-
 478 havior as a heuristic to deal with compound
 479

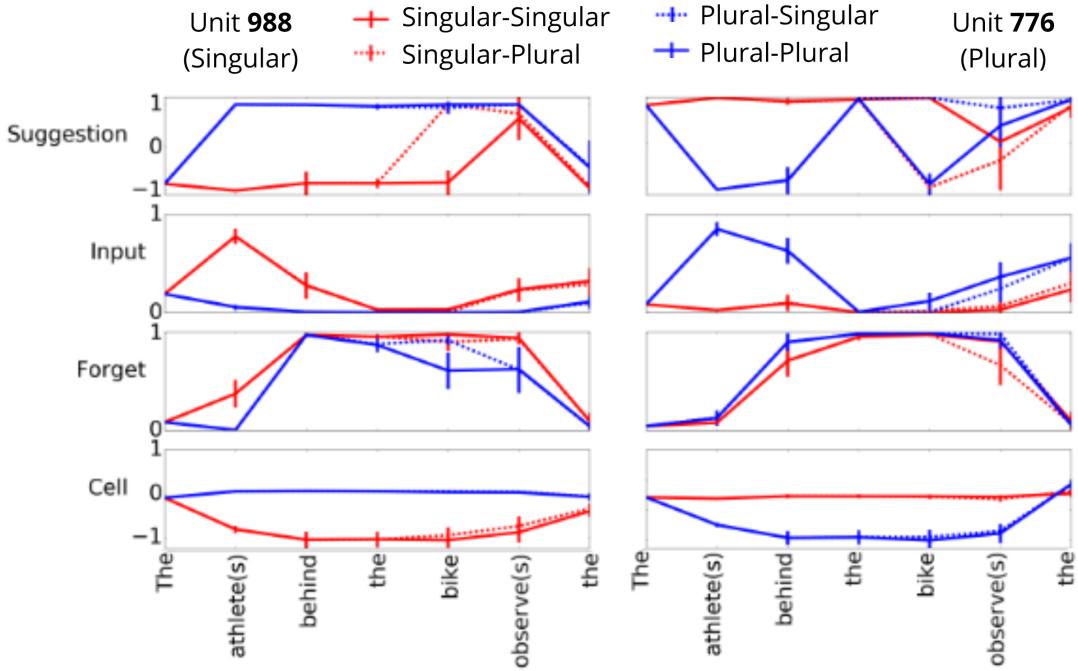
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575

Figure 2: Cell and gate activations during processing of a sentence with a prepositional phrase between subject and verb. (A) Cell activity C_t for the two number units 775 and 987 and output activity h_t for the syntax unit 1149, for all four combinations of grammatical numbers of the two nouns. Note that the cell activity of units 775/987 is non-zero only when the first noun is plural/singular, respectively. (B) Corresponding forget-gate activity for the same number units. Note that gate activity is indifferent of the grammatical number of both nouns and that its value is close to one during the PP until after the verb. (C) Input-gate activity of the same units. Note that the gate value of unit 775/987 spikes around the first noun only when it is plural/singular.

nouns, given that for compound nouns the relevant number information resides at the second noun, whereas in the case of simple nouns there's anyway no risk of encountering an interfering noun immediately after the subject (TODO: discuss this part in the meeting to see if it makes sense to all. If yes, perhaps we could easily check this in an experiment.). Finally, note that for unit 988, the input gate is only open when the subject is singular, whereas for unit 776 it is only open when the subject is plural. This too is consistent with the labeling of these units as singular and plural.

Third, forget-gate dynamics of both number units also seem to correspond to the above properties. In both units, forget-gate activity starts at value around zero $f_{t_{subject}=0}$ and then goes abruptly towards its maximal value at the next

step $t_{subject} + 1$, then stably staying at this level until after the verb $t_{verb} + 1$. Note that for all four conditions (SS, SP, PS and PP), the forget-gate exhibits similar dynamics, being indifferent to the grammatical number of the subject. This seems appropriate for whether the second noun is singular or plural given that the network cannot know in advance whether an interfering noun will appear, and it should anyway store number information for long-range dependencies also in the absence of any upcoming noun (TODO: explain or leave as an open question the reason for which we observe the same dynamics whether the first noun is singular or plural). Last, we note that in all cases the forget-gate activity resets at $t_{verb} + 1$. This seems appropriate, given that at this point the subject's number is no more useful, and the cell would be better free

576 up to encode new number information.
 577 Finally, cell activity should reflect the dynamics
 578 of the suggestion, input and forget gates. Indeed,
 579 the cell value becomes non-zero at $t_{subject}$
 580 and preserves this value until $t_{verb}-1$ when verb-
 581 form prediction occurs (Panels G-H). Note that
 582 this is the case only for the relevant conditions:
 583 in conditions SS and SP, unit 987 encodes singular
 584 as $C_t = -1$ and is approximately zero during
 585 sentence processing in the other two conditions
 586 (PP and PS). Similarly, unit 776 encodes plural
 587 with a non-zero, negative, value only in the
 588 relevant conditions (PP and PS) but not in the
 589 irrelevant ones (SS and SP). Note that for the ir-
 590 relevant conditions, cell activity is kept approx-
 591 imately zero thanks to the clear-up of the cell:
 592 $f_{t_{subject}} = 0$ and $i_{t_{subject}} = 0$, and the following
 593 input- and forget-gate dynamics.
 594 Taken together, these results describe the
 595 intricate mechanism underlying subject-verb
 596 agreement in LSTM number units. They also
 597 clarify why ablating either one of these two units
 598 may bring the network close to chance level on
 599 the NA-task. Without the stored information in
 600 the cell of a number unit the network hopelessly
 601 tries to solve the task.

602

603 5.1.3 Predicting the verb form

604 For long-range dependencies, storing the num-
 605 ber of the subject in the cell is necessary for
 606 correct prediction in the NA-task (except for
 607 cases of mere chance) but is not sufficient. Two
 608 more conditions need to be met: (1) the stored
 609 number should be output from the cell at the
 610 right moment $t_{verb}-1$; and (2) The output from
 611 the number unit should increase activations only
 612 in units that represent the matching verb form
 613 in the output layer. For example, the output
 614 from unit 988 should be projected differently
 615 to singular and plural verb forms in the out-
 616 put layer, such that it will increase activity only
 617 in units representing the singular form. Given
 618 that the encoding of singular by unit 988 is with
 619 a negative value $C_{t_{verb}-1} < -1$, and therefore
 620 $h_{t_{verb}-1} < -1$ (equation 1.x), this means that
 621 weights from unit 988 to units in the output
 622 layer that correspond to singular forms should
 623 be negative too, but not those projecting to plu-

ral forms.
 624 To see whether number units meet the first
 625 condition, figure 3A shows the output-gate and
 626 hidden state dynamics of units 988 and 776. In-
 627 deed, the output gate opens at $t_{verb}-1$, reaching
 628 its almost maximal value (Panels A-B). This en-
 629 sures that the stored number information is out-
 630 put from the cell and propagates to the output
 631 layer via h_t (equation 1.X) at the right moment
 632 (Panels C-D). Note also that for both units,
 633 output-gate dynamics are quite similar across
 634 all four conditions. This may seem sub-optimal,
 635 since one may expect that the network would
 636 learn to close the output gate for the irre-
 637 relevant conditions (PS and PP for unit 987 and
 638 SS and SP for unit 776). However, as we saw,
 639 the cell value for these condition is anyway ap-
 640 proximately zero and therefore an open output
 641 gate will have the same effect as closed one.
 642 Next, to see whether number units meet the
 643 second condition, figure 3B presents the distri-
 644 bution of weight values from the two number
 645 units and from several other units to 36 units
 646 at the output layer - 18 corresponds to verbs in
 647 the singular form and 18 to verbs in the plural
 648 forms (TODO: extend to more/all verbs in the
 649 vocab). Clearly, for number units, weights to
 650 singular and plural forms have different values,
 651 but for other non-number units, there's no clear
 652 structure. Moreover, the weight values corre-
 653 spond to the encoding of singular/plural in the
 654 number unit. For example, weight values from
 655 unit 988 to singular forms are indeed negative,
 656 and those to plural forms are not. For all other
 657 units as well, the sign of $h_{t_{verb}-1}$ corresponds to
 658 the sign of the relevant weight values (panels A-
 659 B), such that their product is always positive.
 660 This ensures that number units increase activa-
 661 tions only in the matching units in the output
 662 layer.
 663

664 5.1.4 Short-range number units

665 We saw in section 5.3.1 that performance on
 666 several NA-tasks was not impaired after the
 667 ablation of any unit, nor by ablating the LR-
 668 number units 776 and 988, which suggested that
 669 number information may be encoded also else-
 670 where in the network, in other units. We there-
 671

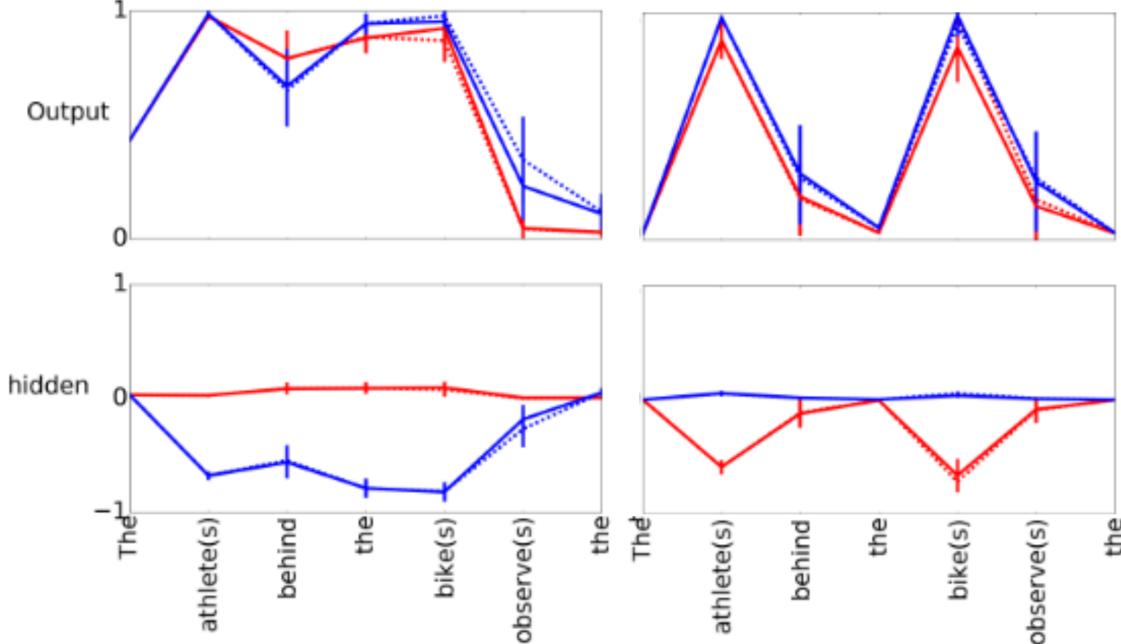
672
673
674
675
676
677720
721
722
723
724
725
726Unit 988
(Singular)+ Singular-Singular
... Singular-Plural+ Plural-Singular
+ Plural-PluralUnit 776
(Plural)

Figure 3: Hidden and output-gate activations during processing of a sentence with a prepositional phrase between subject and verb.

700

748

fore start by defining requirements that a unit should meet to count as a (grammatical) *number unit*. For convenience, we focus on sentences that start by a subject and define two such requirements:

749
750
751
752
753

1. *Continuous code*: The grammatical-number of the subject should be decodable from unit's cell activity in a continuous manner. Continuity means that decoding performance is high for *all* subsequent time points and that the code for grammatical number is the same code as that used by the unit during the presentation of the subject. This excludes units for which grammatical code can be decoded from their activity at subsequent time points but that there is no consistent code, e.g., cell values for singular and plurl flips from

one time point to another. Units that meet this requirement for continuous code could be therefore interpreted as *storing* grammatical-number information.

754
755
756
757
758

2. *Number-segregated efferent weights*: The efferent weights of a number unit should be segrarated for singular and plural forms of the verb. This is similarly to what was described above for the LR number units.

759
760
761
762
763
764
765
766
767

To quantify the first requirement, we use a generalization across time (GAT) approach (cite). Specifically, for each unit, we trained a model to predict the grammatical number of the subject based on cell activity during the presentation of the subject but tested its prediction based on activity from all time points. We evaluate model performance using the Area under of

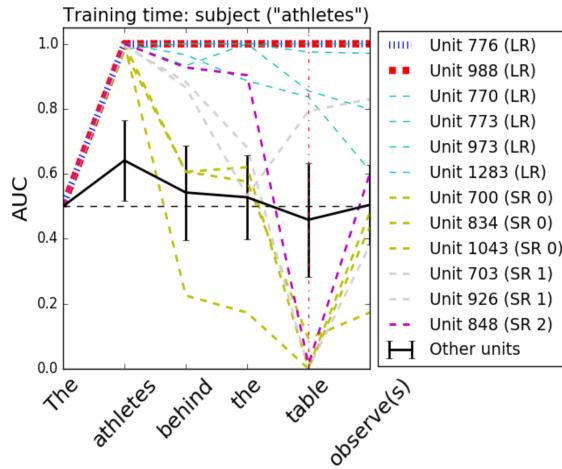
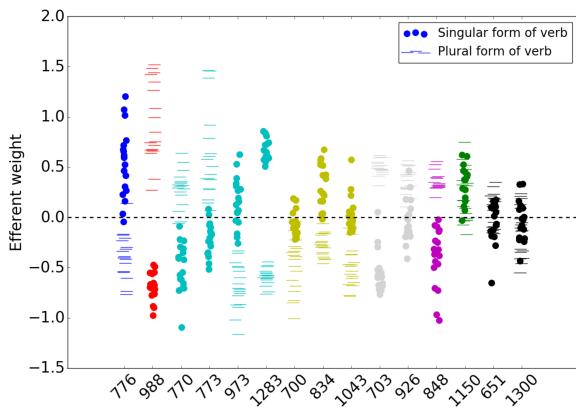
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782

Figure 4: Connectivity structure to output layer. (A) Output activity h_t of all number units during the processing of a sentence with a PP between subject and verb. (B) Weight values from various units to output layer. Note that only for number units the output weights are clearly separated between singular and plural form of the verb, either positive or negative, compare to the syntax unit (1149) and two non-number units in the second layer. (C) Visualization of 18 verbs in their plural and singular forms (36 words in total) on the plane spanned by the two first principal components of their embeddings by the output weight matrix. A clear separation is observed between the singular and plural form along the first PC.

Curve (AUC), and for each time point, we define 'high' performance as AUC values greater than 0.95. To quantify the second requirement, given two sets of weights to singular and plural verb forms W_S, W_P of unit u , with means \bar{W}_S, \bar{W}_P and standard deviations σ_S, σ_P , we define a statistical distance: $d_{W_P, W_S}^u = \frac{|\bar{W}_S - \bar{W}_P|}{\sigma_S + \sigma_P}$. A unit is then said to have highly segregated weights if d_{W_P, W_S}^u is at the top five percentile across all units. Units that meetFinally, as discussed in section 5.3.1, we distinguish between two types of units: (1) SR-units - units

Figure 4B presents shows the Area Under of Curve (AUC) for all units in the network that meet these two requirements.

5.2 Syntax units

[1]

5.2.1 Predicting syntactic-tree depth from network activity

[1]

5.2.2 Ablation study

[1]

5.3 Syntax-number units interactions

[t] [1]

5.4 Processing of relative clauses

[1]

[1]

6 Discussion

[1] Bock + asymmetry plural/singular (non-phonological explanation)

Acknowledgments

[1]

816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838

839

840

841

842

843

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863

864	912
865	913
866	914
867	915
868	916
869	917
870	918
871	919
872	920
873	921
874	922
875	923
876	924
877	925
878	926
879	927
880	928
881	929
882	930
883	931
884	932
885	933
886	934
887	935
888	936
889	937
890	938
891	939
892	940
893	941
894	942
895	943
896	944
897	945
898	946
899	947
900	948
901	949
902	950
903	951
904	952
905	953
906	954
907	955
908	956
909	957
910	958
911	959

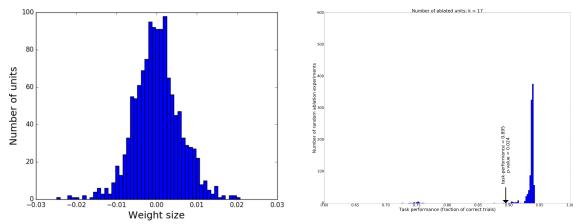
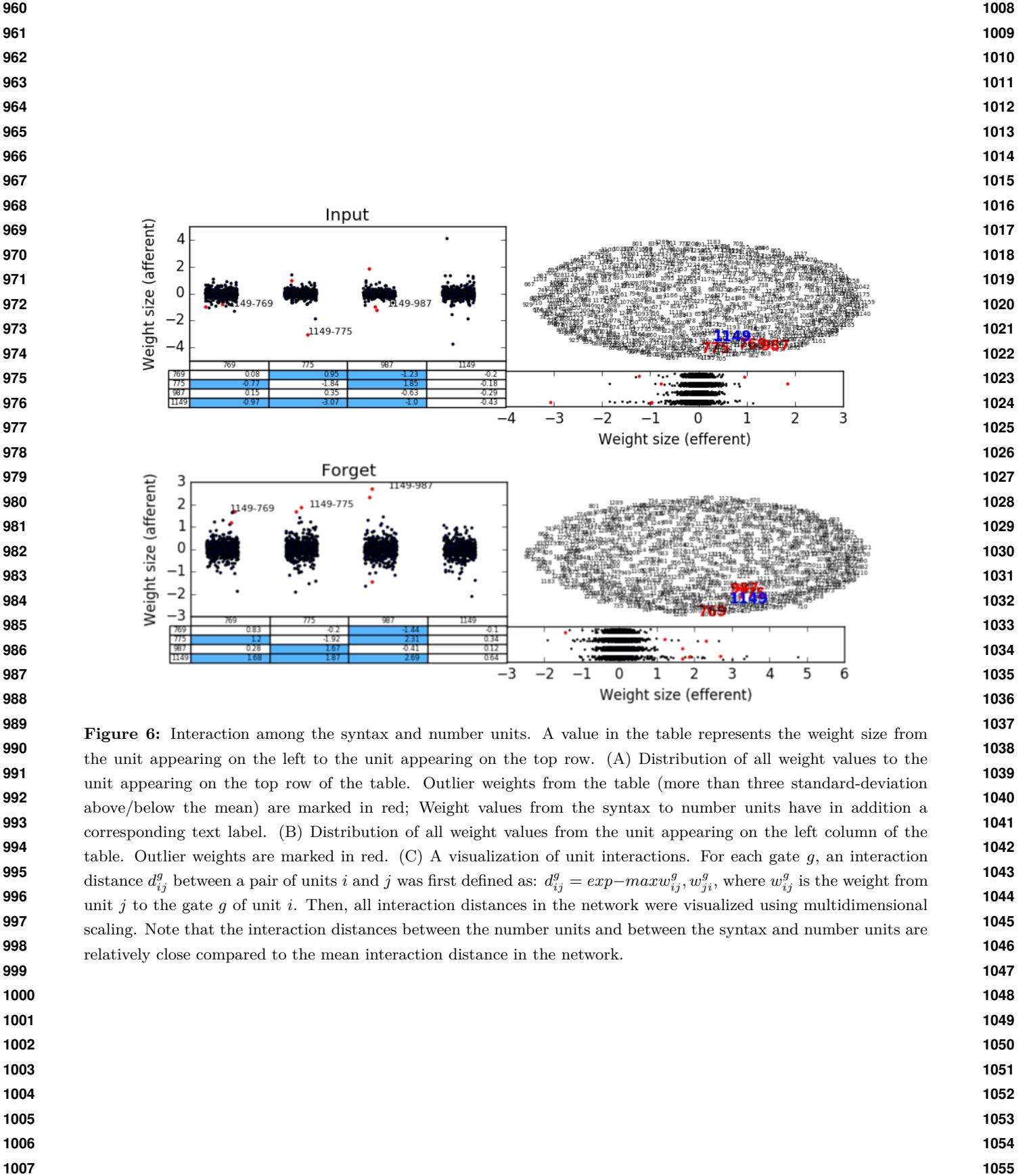


Figure 5: (A) Distribution of the resulting weight values from the tree-depth regression model. Outlier weights were defined as having a value that is distant from the mean by more than three standard deviations (17 outlier weights in total - marked in red). (B) Task performance of 1000 models after ablating 17 random units (in blue) and based on the 17 outlier weights from the tree-depth regression model (black arrow). The reduction in performance due to outlier-weights ablation is statistically significant ($p - value < 0.05$) when compared to the null distribution generated by the random ablations.



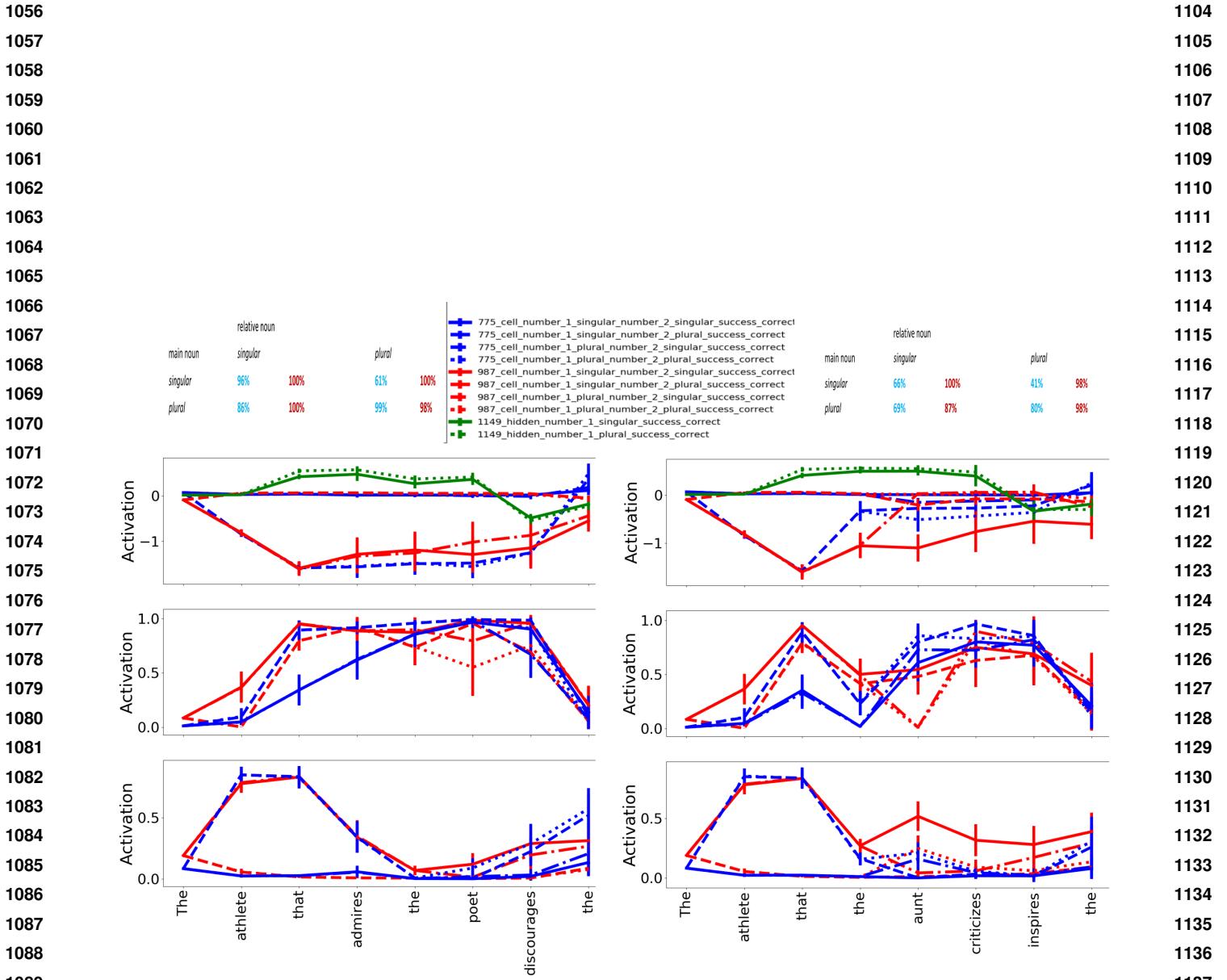


Figure 7: Subject-verb agreement in relative clauses: agreement-task accuracy for (A) subject relatives and (B) object relatives. (C & D) The corresponding cell activations for the number units (775 and 987) and the syntax unit 1149. (E & F) The corresponding forget-gate activity and (G & H) input-gate activity of the number units.

