Sparse Encoding of Grammatical Gender in LSTM Language Models

Priyanka Sukumaran (psukumaran23@gmail.com)

Conor Houghton (conor.houghton@bristol.ac.uk)
School of Engineering Mathematics and Technology, University of Bristol

Nina Kazanina (nina.kazanina@unige.ch)
Department of Basic Neurosciences, University of Geneva

Abstract

Neural network language models excel at capturing the complexities of natural language, yet their internal representations remain poorly understood. A key question is whether such models form structured, human-like abstractions that support generalization. We investigate how LSTM language models encode grammatical gender—an ideal test case, as gender is lexically fixed and generally not inferable from semantics. We focus on long-distance dependencies and various gender agreement configurations.

We conduct single-unit ablation to identify neurons critical for grammatical gender agreement. Across eight LSTM models, we find between one and five units whose removal significantly disrupts performance—by over 40% in some constructions involving gender-interfering nouns. These units are essential for both noun-adjective and noun-past-participle gender agreement. Neuron activity analyses reveal that these units exhibit category-specific effects, with some showing a preference for default gender forms, such as masculine nouns.

Our findings show that LSTMs develop sparse and structured representations of grammatical gender, reminiscent of grandmother cells in neuroscience. These results suggest that abstract grammatical categories can emerge naturally in LSTM training. More broadly, this work contributes to our understanding of how language models encode linguistic structure, with implications for model interpretability and parallels between artificial and biological computation.

Keywords: neural network language models; linguistic generalization; mechanistic interpretability; encoding mechanisms

Introduction

LANGUAGE, a uniquely human cognitive function, is now remarkably imitated by neural network language models. Language models perform well across a range of language comprehension and generation benchmarks, sometimes reaching human-level performance. Yet, their errors remain revealing: where they fail, we gain insight into the distinctiveness of human linguistic competence, offering valuable comparisons for cognitive science (Chowdhury & Zamparelli, 2018; Chaves, 2020; Lan et al., 2024). Moreover, understanding how language models perform linguistic tasks offers cognitive science new comparative tools for studying the neural basis of language. In particular, the implicit grammatical representations in machine models may provide testable hypotheses about how grammar is represented neurally (Houghton et al., 2023).

This study focuses on LSTMs to investigate the mechanisms of grammatical processing using the targeted syntactic evaluation approach Linzen et al. (2016). LSTMs demonstrate strong grammatical competence and greater biological plausibility than more powerful transformer models, positioning them as effective minimal models for studying neural grammar representation (Linzen et al., 2016). We investigate the internal mechanisms that enable LSTMs to encode and represent grammatical gender to maintain agreement in complex contexts. We use single-neuron ablation: systematically removing or 'ablating' neurons from a neural network to understand their function. In addition, we ask whether LSTM language models encode the grammatical gender as an *abstract category*, thus enabling linguistic generalization.

This paper confirms the emergence of specialized single neurons and sparse networks in LSTMs for encoding grammatical gender, using a test set we developed in French. We investigate two gender agreement contexts: noun-adjective and noun-past-participle. We evaluate the role of gender-specific units in these contexts. For each, we test both adjacent and long-distance agreement in the presence of attractors. Expanding on (Lakretz et al., 2019), which focused on constructions with minimal intervening words, we evaluate individual neurons' roles across varying dependency lengths. We assess whether LSTMs generalize grammatical gender across four head noun categories: singular/plural and feminine/masculine. Our findings provide evidence of sparse and selective representations in LSTMs, highlighting parallels with biological systems such as the *grandmother cell* phenomenon in neuroscience (Gross, 2002; Quiroga et al., 2005).

Methods

We trained a two-layer LSTM with 650 hidden units on French Wikipedia data (Mueller et al., 2020) for next-word prediction. We trained both tied and untied versions of the model, with the tied variant sharing input, output, and embedding weights (Press & Wolf, 2017) to facilitate interpretability. For robustness, we trained five untied and three tied models with different random initialization seeds.

The LSTM models were tested on grammatical agreement, which involves coordinating the noun with other elements such as the verb, determiner or adjective based on the noun's properties such as nominal number (singular/plural) and gender (masculine/feminine/neutral). Traditional psycholinguis-

tics studies have used agreement tasks to probe hierarchical syntactic knowledge that humans employ to parse language (Bock & Miller, 1991; Franck et al., 2002).

We evaluate LSTM performance before and after single-neuron ablation for two common contexts of gender agreement in French: noun-adjective (NA) and noun-participle (NP) agreement, see Table 2. The simplest case is (**A**) adjacent agreement, where no intervening words separate the noun and its agreement target, making agreement straightforward. For instance, in the NA construction in Table 2, $\text{la}_{\text{s.f}}$ **robe$_{\text{s.f}}$** $\underline{\text{est}}_{\text{s.f}}$. To systematically test agreement, we varied the gender and number of the subject noun. Secondly, we also test (**B**) long-distance agreement by including 1-11 words between the noun and its agreement target using prepositional phrases and subject-relative clauses. These constructions are labelled NA-n and NP-n, where n is the number of intervening words as an example for NA-5: la **robe$_{\text{s.f}}$** $[\text{que j' aime beaucoup}]_{\text{rel}}$ $\underline{\text{est}}_{\text{s.f}}/\text{bleu}_{\text{m.f}}$. Next, we test the more complex (**C**) agreement across an attractor by introducing another noun, with varying number and gender, using a prepositional phrase that could potentially interfere with gender agreement. We test this using conditions **NN$_{\text{num.gen}}$A** and **NN$_{\text{num.gen}}$P**, see Table 2.

Table 1: Gender/syntax units in different models with a random seed, labeled as M_{seed}. This result section focuses on M_{528}. Each unit reduces gender agreement performance when: SG: the head noun is singular, PG: the head noun is plural, G: the head noun is either singular or plural, S: there is an interfering attractor or in longer dependencies.

	Gender/syntax units			
Untied Model	**SG**	**PG**	**G**	**S**
M_{528}	-	870	1098	958
M_{1571}	914	1013	-	-
M_{704}	863	1012	-	-
M_{220}	-	-	937	-
M_{73}	-	-	1269	-
Tied Model				
M_tied_{528}	-	-	1174	-
M_tied_{704}	-	1262	930	
M_tied_{73}	764	768	-	-

Results & Discussion

Across eight LSTM initialisations, ablating just 2–3 out of 1300 units resulted in significant drops in grammatical agreement performance. We focus on model M_{528}, with similar patterns observed across other runs (Table 1). Three units consistently emerged as critical: Unit 870 (**PG-870**), which selectively encoded plural gender agreement; Unit 1098 (**G-1098**), encoding gender more broadly; and Unit 958 (**S-958**), which tracked syntactic structure. Performance on long-distance dependencies—especially those with intervening tokens or attractors—dropped to near chance after ablation.

The single-unit gate and cell state activity analyses confirmed that these units encoded abstract grammatical features. Gender units activated at the head noun and maintained gender information through to the agreement target, often via sustained cell state values. Syntax units preserved structural information across the clause, suggesting internal representations of sentence depth. t-SNE projections of unit activations revealed clear separability by gender and number, further supporting the presence of abstract, category-sensitive encodings (Figures 7 and 8).

Interestingly, the effect of ablation was asymmetric: agreement with feminine nouns was far more disrupted than masculine, aligning with the default reasoning strategy proposed by (Jumelet et al., 2019), whereby default grammatical features—such as singular number or masculine gender—are encoded implicitly in model weights, while non-default features like plural or feminine depend more on explicit, input-driven encoding. Our results support this distinction: ablations of gender-selective units impaired feminine agreement but left masculine largely intact, indicating that masculine may be redundantly or diffusely encoded.

These findings mirror a broader distinction in neural coding between localist and distributed representations. The emergence of highly selective units in our LSTM aligns with the grandmother cell hypothesis in neuroscience—the idea that individual neurons (or units) can become tuned to specific, abstract categories. Notable examples include the "Jennifer Aniston neuron" in the human hippocampus, which responds exclusively to images of that individual (Quiroga et al., 2005). Similarly, Konorski (1967) proposed gnostic units for recognizing object categories. Our gender and syntax units exhibit similar behavior in a computational setting, selectively activating for abstract grammatical roles and sustaining them across syntactic contexts.

However, localist units do not exclude distributed coding. Sparse, localist encodings (as in our identified units) can coexist with more distributed representations—particularly for default categories. This balance reflects ongoing debates in neuroscience: while localist encoding offers interpretability, distributed representations are thought to be more robust and biologically plausible (Bowers, 2017; Rolls, 2017). Importantly, distributed representations can also be sparse—involving only a few active units, each of which may not encode easily interpretable features.

Thus, LSTMs, while simplistic compared to the brain, offer a computational testbed for examining these dual encoding strategies. These findings open several avenues for future work. Do the identified units also encode related features like animacy, case, or number? Are they robust across languages with richer agreement systems (e.g., German, Bantu)? Can similar sparse mechanisms be observed in transformer models, or are they unique to sequential architectures like LSTMs? Most importantly, ablation studies like ours help demarcate the boundary between what linguistic phenomena can be explained by statistical learning in neural networks, and what might be uniquely human. As noted by Surendra et al. (2023), this boundary is often subtle.

References

Bock, K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, *23*(1). doi: 10.1016/0010-0285(91)90003-7

Bowers, J. S. (2017). *Grandmother cells and localist representations: a review of current thinking* (Vol. 32) (No. 3). doi: 10.1080/23273798.2016.1267782

Chaves, R. P. (2020). What don't RNN language models learn about filler-gap dependencies? . In *Proceedings of the society for computation in linguistics.*

Chowdhury, S. A., & Zamparelli, R. (2018). RNN simulations of grammaticality judgments on long-distance dependencies. In *Coling 2018 - 27th international conference on computational linguistics, proceedings.*

Franck, J., Vigliocco, G., & Nicol, J. (2002). Subject-verb agreement errors in French and English: The role of syntactic hierarchy. *Language and Cognitive Processes*, *17*(4). doi: 10.1080/01690960143000254

Gross, C. G. (2002). *Genealogy of the "grandmother cell"* (Vol. 8) (No. 5). doi: 10.1177/107385802237175

Houghton, C., Kazanina, N., & Sukumaran, P. (2023, 12). Beyond the limitations of any imaginable mechanism: Large language models and psycholinguistics. *Behavioral and Brain Sciences*, *46*, e395. doi: 10.1017/S0140525X23001693

Jumelet, J., Zuidema, W., & Hupkes, D. (2019). Analysing neural language models: Contextual decomposition reveals default reasoning in number and gender assignment. In *Conll 2019 - 23rd conference on computational natural language learning, proceedings of the conference.* doi: 10.18653/v1/k19-1001

Konorski, J. (1967). *Integrative activity of the brain*. Chicago: University of Chicago Press.

Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., & Baroni, M. (2019). The emergence of number and syntax units in LSTM language models. In *Naacl hlt 2019 - 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies - proceedings of the conference* (Vol. 1).

Lan, N., Chemla, E., & Katzir, R. (2024). Large language models and the argument from the poverty of the stimulus. *Linguistic Inquiry*.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, *4*. doi: 10.1162/tacl{_}a{_}00115

Mueller, A., Nicolai, G., Petrou-Zeniou, P., Talmina, N., & Linzen, T. (2020). Cross-Linguistic Syntactic Evaluation of Word Prediction Models. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5523–5539). doi: 10.18653/v1/2020.acl-main.490

Press, O., & Wolf, L. (2017). Using the output embedding to improve language models. In *15th conference of the european chapter of the association for computational linguistics, eacl 2017 - proceedings of conference* (Vol. 2). doi: 10.18653/v1/e17-2025

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*(7045). doi: 10.1038/nature03687

Rolls, E. T. (2017). *Cortical coding* (Vol. 32) (No. 3). doi: 10.1080/23273798.2016.1203443

Surendra, K., Schilling, A., Stoewer, P., Maier, A., & Krauss, P. (2023). Word class representations spontaneously emerge in a deep neural network trained on next word prediction. In *2023 international conference on machine learning and applications (icmla)* (pp. 1481–1486).

Table 2: Example phrases for adjacent noun-adjective and noun-past-participle gender agreement. We tested both singular and plural noun phrases for each condition. s.f: singular feminine, s.m: singular masculine, p.f: plural feminine, p.m: plural masculine.

	Singular	**Plural**
Noun-adjective		
NA *(No attractor)*	la **robe** est bleue/bleu the **dress**$_{s.f}$ is blue$_{s.f}$/blue$_{s.m}$	les **robes** sont bleues/bleus the **dresses**$_{p.f}$ are blue$_{p.f}$/blue$_{p.m}$
NN$_{gen}$A *(Gender attractor)*	la **robe** avec le *sac* est bleue/bleu the **dress**$_{s.f}$ with the *bag*$_{s.m}$ is blue$_{s.f}$/blue$_{s.m}$	les **robes** avec les *sacs* sont bleues/bleus the **dresses**$_{p.f}$ with the *bags*$_{p.m}$ are blue$_{p.f}$/blue$_{p.m}$
NN$_{num.gen}$A *(Number/gender attractor)*	la **robe** avec les *sacs* est bleue/bleu the **dress**$_{s.f}$ with the *bags*$_{p.m}$ is blue$_{s.f}$/blue$_{s.m}$	les **robes** avec le sac sont bleues/bleus the **dresses**$_{p.f}$ with the *bag*$_{s.m}$ are blue$_{p.f}$/blue$_{p.m}$
Noun-participle		
NP *(No attractor)*	la **robe** est tombée/tombé the **dress**$_{s.f}$ fell$_{s.f}$/fell$_{s.m}$	les **robes** sont tombées/tombés the **dresses**$_{p.f}$ fell$_{p.f}$/fell$_{p.m}$
NN$_{gen}$P *(Gender attractor)*	la **robe** avec le *sac* est tombée/tombé the **dress**$_{s.f}$ with the *bag*$_{s.m}$ fell$_{s.f}$/fell$_{s.m}$	les **robes** avec les *sacs* sont tombées/tombés the **dresses**$_{p.f}$ with the *bags*$_{p.m}$ fell$_{p.f}$/fell$_{p.m}$
NN$_{num.gen}$P *(Number/gender attractor)*	la **robe** avec les *sacs* est tombée/tombé the **dress**$_{s.f}$ with the *bags*$_{p.m}$ fell$_{s.f}$/fell$_{s.m}$	les **robes** avec le *sac* sont tombées/tombés the **dresses**$_{p.f}$ with the *bag*$_{s.m}$ fell$_{p.f}$/fell$_{p.m}$

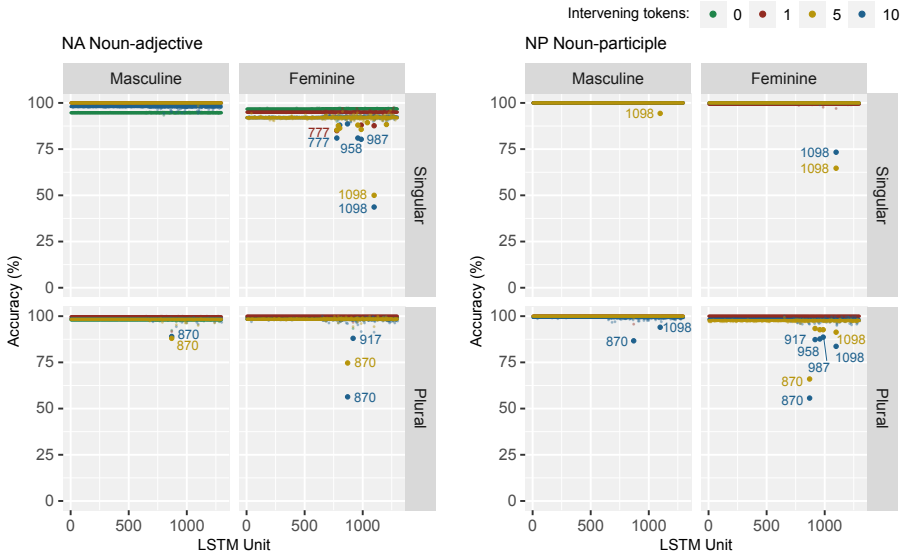

Figure 1: Performance on simple agreement with 0, 1, 5 and 10 intervening tokens for noun-adjective (NA) and noun-participle (NP) conditions after ablation of each LSTM unit (x-axis). Each panel shows mean agreement accuracy (y-axis) after ablating individual units (x-axis) for singular/plural and masculine/feminine nouns. Performance is split by head noun category: singular (top), plural (bottom), masculine (left) or feminine (right). Performance is further broken down by constructions with zero (green), one (red),five (yellow), and ten (blue) intervening tokens. Each dot represents performance after ablating a unit, with significant drops (z-score < -3) highlighted.

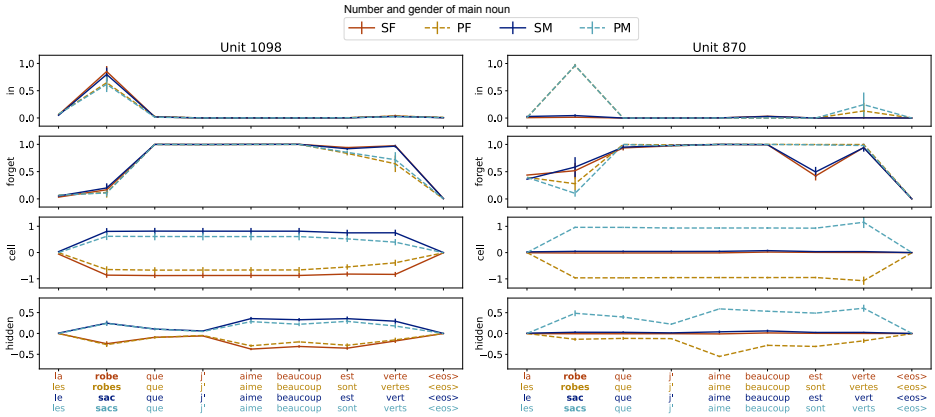

Figure 2: Input gate, forget gate, cell state and hidden state activity for gender units for simple agreement, that is, no attractor.

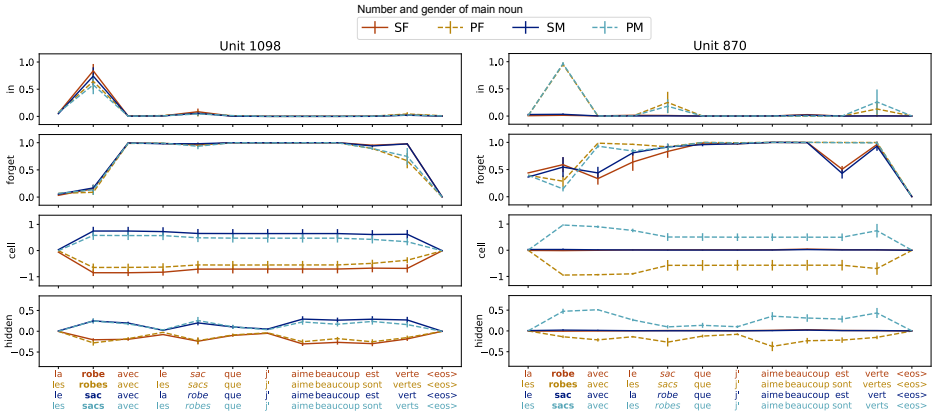

Figure 3: Input gate, forget gate, cell and hidden state activity for agreement across a gender attractor ($NN_{gen}A$/ $NN_{gen}P$)

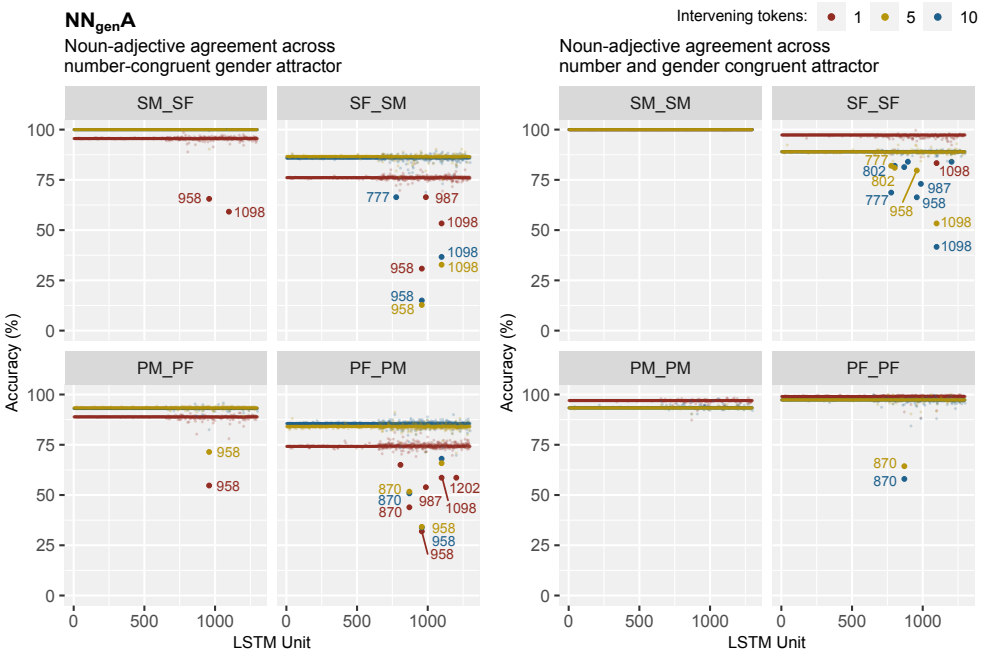

Figure 4: Performance after ablation of each LSTM unit (x-axis) on noun-adjective agreement with attractor nouns of varying gender. Each panel corresponds to test sentences with different combinations of subject and attractor noun categories. For example, the SM_SF indicates Singular-Masculine subject noun and Singular-Feminine attractor.

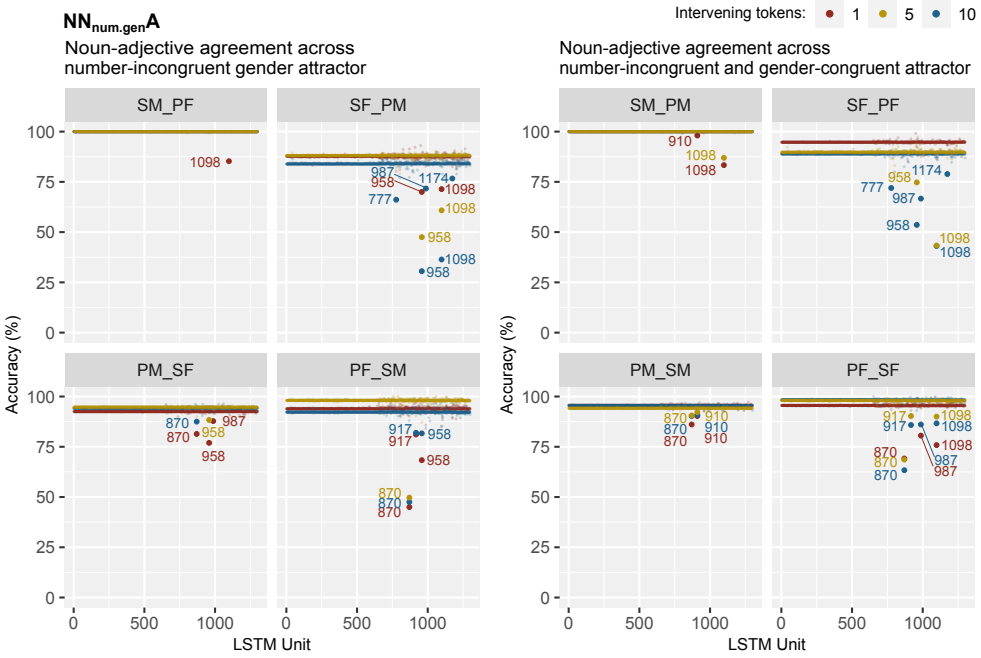

Figure 5: Performance after ablation of each LSTM unit (x-axis) on noun-adjective agreement with attractor nouns of varying gender and number.

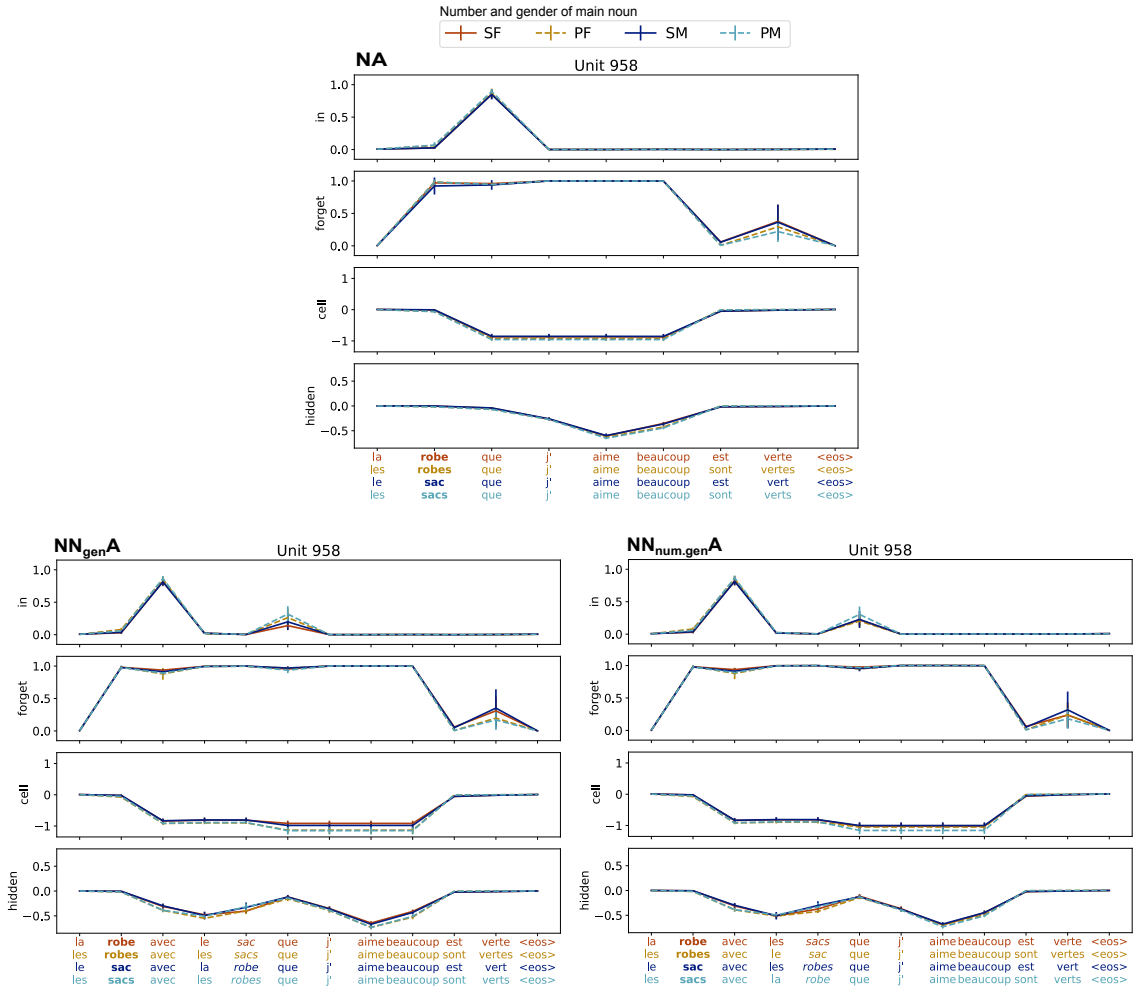

Figure 6: Input gate, forget gate, cell and hidden state activity for syntax unit **S-958** across conditions

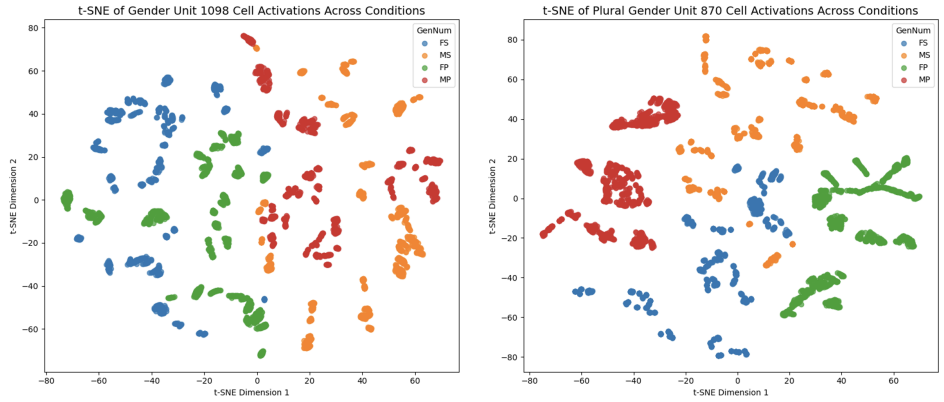

Figure 7: t-SNE projections of LSTM cell activations for selected gender units across gender agreement conditions (FS = feminine singular, FP = feminine plural, MS = masculine singular, MP = masculine plural). Activations across six syntactic configurations (simple agreement, agreement across gender and number attractors), both for noun-adjective and noun-participle across short and longer-range dependencies, were aggregated and flattened across time steps to capture the full temporal profile. Gender **Unit-1098** (left) and Gender **Unit-870** (right) show clear clustering by grammatical gender, suggesting category-specific encoding across sentence constructions. In contrast, syntax unit in Figure 8 shows less separability.

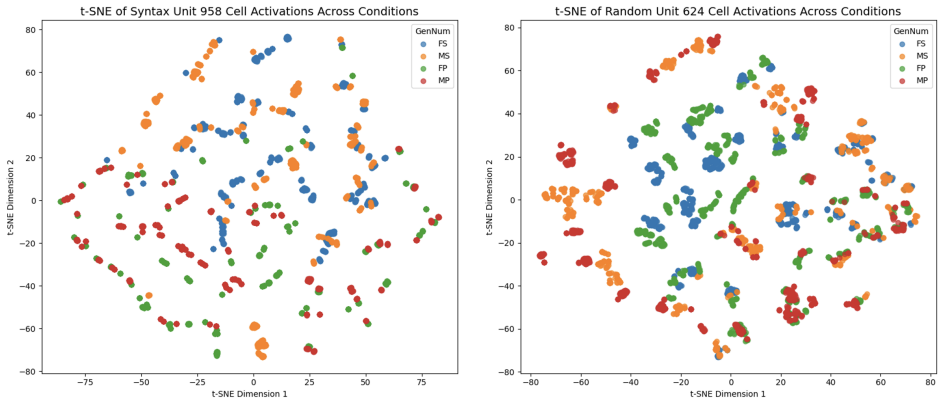

Figure 8: t-SNE projections of LSTM cell activations for the syntax-related **Unit-958** (left) and a randomly selected unit **Unit-624** (right). Compared to the gender units (Figure 7), these units show less separable patterns across gender-number categories, highlighting the contrast between specialized and non-specialized representations.