Corso di Psicolinguistica Computazionale, Università di Pisa
Novembre 2023

Assignment:

A quantitative analysis of the morphological complexity using French verbal system data

Ilaria Salogni

Notebook with code can be found on Github

1. Introduction

As claimed in [1], quantitatively assessing the comparative complexity of inflectional systems across languages is an hard task, because of typological diversity. The processing-oriented approach proposed by the same authors in [2] consists in the training of a Temporal Self-Organizing Map (TSOM, a recurrent variant of Kohonen's Self-Organizing Maps [3]) , on a set of top-frequency verb forms, with no added information about the morphosemantic and morphosyntactic content conveyed by the forms, and assessing the behaviour of the network after the training.

While a TSOM does have the capability to capture and adapt to the temporal context up to the previous time step (*t-1*), it doesn't explicitly keep track of the entire sequence globally across all neurons. Furthermore, the constraints on the topological organization of the map avoid data overfitting, as the map cannot possibly build up a dedicated memory traces for each form in the training set (i.e., it cannot memorize the original input data in its entirety) [4].

In this assignement, I will proceed with the quantitative analysis of the outputs[1] of a TSOM after 100 epochs of training on a corpus of inflected verb forms from 50 top frequency French verb paradigms, discussing in conclusion how morphological inflection can be seen as "*a collective, emergent system, whose global self-organization rests on a surprisingly small handful of language-independent principles of word coactivation and competition*" [1]

2. Dataset

To provide data-driven evidence of how, following a graded-notion of regularity/irregularity, the processing of the morphological structure of a verb could work, 9-14 inflected verb forms have been selected for each of the 50 top frequency French verb paradigms. For this assignment, a dataset resulting from the serial (symbol by symbol) processing of each form by a TSOM for 5 epochs, after 100 epochs of training on the same forms, was provided. The repetition over the five provided epochs was treated as a random effect, assuming no consistent and systematic influence, and I utilized data from all epochs to enhance the statistical robustness by expanding the sample size. For this project, a language not analyzed in the task organizers previous work [1] is observed. The variables contained in the dataset and prominent for the analysis concerned were:

1. **Regularity measures**

Both dichotomous labelling (*IR* in the dataset) and graded regularity annotation were carried out for each paradigm in the dataset.
Irregular paradigms undergo unpredictable stem formation processes: the Levensthein distance (minimum number of single-character edits that are required to transform one string into another, 0 means same string) is computed as *wordNNBstem0*. Then, defining a stem-family as the set of formally distinct stem-sharing members, the average stem-family size was calculated and then normalized by dividing it by the maximum number of possible stem-sharing members in the paradigm. This score, which we will refer to as gradient paradigm regularity (*paradigm.Regularity* in the dataset), ranges between 0 (no paradigm member share the stem) and 1 (all the forms share the same stem, and a paradigm is fully regular). Paradigm regularity score is the same for all members of a paradigm. The two notions are clearly correlated (in our data, p-value < 2.2e-16).

2. **Prediction rate** The prediction rate (*AnticipBMU* counts how many consecutive letters are predicted by the map during word processing, adding +1 point for every correctly predicted symbol (or 0 otherwise).

3. **Entropy** Pointwise entropy (*pointwiseH* in the dataset) was provided for each symbol, calculated using the probability obtained applying the softmax function to the activation value for each unit in the TSOM. I assumed that pointwise entropy was calculated as:

$$H(x) = -log_2 P(x)$$

4. **Probability** As I don't have access to the TSOM, I calculated the probability (*prob* in the dataset) using the formula:

$$P(x) = 2^{-}H(x)$$

This probability measure reflects the likelihood of a particular unit being selected given the input pattern and the weights associated with each unit.

3. Related work

As in Rorberi's master thesis[5], Marzi [6] designed the experiment that is re-proposed in this assignment, focusing on the verbal inflection of Russian to assess the perception of regularity gradients, transcending Pinker and Ulhman[7] model. In Basso's Master thesis[8], a similar TSOM network was trained exclusively with derived forms in Italian, belonging to three inflectional families.

[1] The notebook with the code can be found on Github or at this link

4. How a TSOM works

As shown in the supplementary material attached to [1], A TSOM consists of a D-dimensional input vector, and a map composed by N nodes. The nodes, topologically organized in a two-dimensional Euclidean space, are fully connected to the input vector, through the spatial connection layer and to all the nodes of the map through the temporal connection layer. In other terms, a time series (a word) of a certain number of symbols (letters) is input to a TSOM as a sequence, and letter by letter each symbol at the time t is encoded as an input vector with D components, and propagated through the map.

Very minimally, these are the steps of the computation:

1. Weight Initialization: At the beginning of the training process, the weights of the nodes in the TSOM are initialized.
2. Recoding: When an input pattern is presented to the TSOM at time t, the similarity between the input and the weights of each node is calculated using a distance metric, in this case as Euclidean proximity.
3. Node Selection: The node whose weight vector is most similar to the input pattern is selected as the "winning" node or the Best Matching Unit (BMU). This node is considered to be the best representation of the input at time t.
4. Update Weights: After selecting the BMU, the weights of the BMU and its neighboring nodes are adjusted to better match the input pattern. This adjustment is done to adapt the TSOM to the temporal patterns in the input data.
5. Learning Temporal Patterns: Through repeated presentations of input patterns over time, the TSOM learns to organize its nodes in a way that captures the temporal structure of the input data. Nodes that are frequently selected for similar input patterns become more specialized in representing those patterns.
6. Topological Organization: The topological organization of the TSOM reflects the relationships between nodes. Nodes that are close to each other in the map are expected to respond similarly to similar input patterns. The spatial arrangement of nodes thus represents the inherent temporal relationships in the data.

This brief excursus is aimed at trivially seeing that in a TSOM only the t-1 instant of time is considered, and constraints prevent overfitting, or in simpler words, the map from memorizing the entire original input data, by avoiding the creation of dedicated memory traces for each form in the training set[4]. The output that we observe is then given by way more intricate dynamics than simple corpus-based frequency computation, although frequency plays a role indeed.

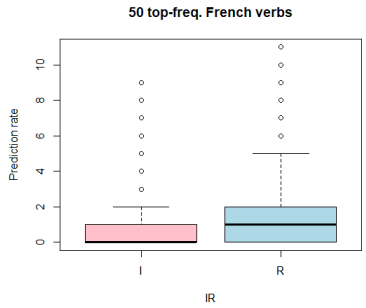

Figure 1: Boxplot distribution of prediction rate through regular and irregular forms

5. Results and discussion

The tests outlined in this assignment were conducted using R on Colab, and detailed procedures and results can be found in the accompanying notebook for clarity and reference.

Regulars vs Irregulars Boxplot distributions (Figure 1) of the prediction rate for the regular and irregular paradigms show that in a very general way, regular verbs are easier to anticipate for the map.

Although as we will see the prediction follows a more constant trend in the irregular shapes, it however shows an overall lower accuracy of the TSOM when compared to the regularly-inflected forms.

Morpheme boundary In the subsequent discussion, a crucial factor to consider while assessing the predictive capability of both regular and irregular verbal forms in this serial task is the position of a character relative to the stem-suffix boundary. After running the Pearson correlation test ($p<2.2e-16$), I depicted (Figure 2) the linear regression model illustrating the prediction rate on a letter-by-letter basis, with the distance from the Morpheme Boundary (MB) as the independent variable (where MB=0 represents the initial element of the suffix). A positive slope signifies that the map becomes increasingly accurate in predicting the upcoming symbol as more of the word input is consumed. Additionally, the consistent slope for both regular and irregular paradigms indicates an absence of discernible interaction between regularity and distance to MB. Once again, analogous to the preceding boxplot analysis, the Temporal Self-Organizing Map (TSOM) demonstrates greater accuracy in predicting regular forms.

Introducing non-linearity Using a non-linear regression model unveils a more complex phenomenon,

where the prediction rate experiences a sharp decline at the stem-suffix boundary. This occurrence is interpreted in [4] as an effect of structural discontinuity between the stem and the suffix. This discontinuity is more pronounced in regularly inflected forms than in irregular ones, given that irregular forms typically exhibit earlier discriminability.

Discriminability Discriminability is how easily an inflected form is discriminated from its paradigm members: for example, the Complex Uniqueness Point (CUP) of *suis* is earlier than the CUP for *manger*, that can only be found at the suffix's beginning. However related to how recurrent a sequence is, it is a metric of how confusable an input sequence is with other.
Irregular items are discerned more easily than regularly inflected ones, resulting in a smoother growth pattern, as depicted in Figure 3, rather than the peaked pattern observed in regular forms. This smoother growth arises from the recurrence of the stem, which initially facilitates processing but becomes more challenging due to the greater diversification (and less reiteration) of suffixes. As elaborated in Marzi (2019), the initial processing disadvantage in irregular paradigms is offset by a reduced effort in transitioning from the stem to the subsequent suffix. Conversely, the processing facilitation in regularly inflected forms is counteracted by the uncertainty incurred at the Morpheme Boundary (MB). The prediction rate sharply declines at the stem-suffix boundary, indicating that when the last letter of a stem is identified, the map revises its expectations for the next symbol. This is interpreted in [1] as a result of a more profound structural discontinuity between the stem and the subsequent suffix in regularly inflected forms compared to irregularly inflected forms. This effect can be seen across at least the 6 languages (Arabic, English, German, Greek, Italian and Spanish) taken into consideration in [1], as well as of course French.

Cross-lingual perspective Similarly, in the work of Rorberi [5], non-linear regression plots depicting interaction effects between morphological (ir)regularity and distance to the Morpheme Boundary (MB) on a Russian dataset structured analogously to the one employed in this assignment, exhibit a consistent pattern. This observation underscores that, despite cross-linguistic variations in the depth of the fall or the steepness of the rise, influenced by factors such as the number of inflection markers and their formal distinctiveness, as discussed in [1], there exists a notable similarity in the overall shape of the curves across different languages.

Gradual ir(regularity) Finally, in Figure 4 the regression plots of the interaction between prediction rate and morpheme bound using 4 different values of ir(regularity), calculated introducing 4 ranges in *paradigm.Regularity*, it can be seen that a gradual vision of verbal regularity also reflects the gradual change of the plot shape.
Reiteration Inflectional systems comprise big families of stem-sharing forms, categorized into regularly inflected (regular forms) and less predictable verb forms (irregular forms). Paradigm regularity facilitates stem processing in the TSOM [1] due to the reiteration of redundant stems in regular forms, leading to a significant reduction in processing prediction at the stem-suffix boundary.

High-frequency and entrenched chains
Furthermore, high-frequency forms tend to be isolated within their own paradigm, limiting the spread of morphological information. As highlighted in [4], In a TSOM, this results in biased expectations, as high-frequency words develop entrenched node chains, favoring a few endings, so the map's propensity to acquire low-frequency and novel endings diminishes. Deeply entrenched node chains within paradigms share little information with other chains, hindering words from different paradigms from benefiting through shared connections. Conversely, uniform frequency distributions within highly entropic paradigms enhance accessibility, favoring global (as in paradigm-based) acquisition.

Entropy The collapse of the processing advantage accumulated across regular stems by suffix selection can be seen as a greater uncertainty of the network, which is confirmed in the values of the punctual entropy of each symbol. This is confirmed by the non-linear regression plot in Figure 5. From the theoretical framework that we have described so far, we also expect the prediction capacity of the network (as described in 2.2) to be inversely proportional to the pointwise entropy. I then set up a linear model, using lm function in R in the notebook. The multiple R-squared value indicates that the model accounts for about 29.15% of the variability in the response variable, and the F-statistic assesses the overall significance of the model. The low p-value (< 2.2e-16) suggests that the model is highly significant, and the Q-Q plot of the model can be found in the notebook.

Finally I also plotted the probability (as calculated in 2.2) for each position against the MB of the correct prediction of the current symbol by the TSOM (Figure 6).

GAM Like originally proposed in [1], I replicated a Generalized Additive Model (GAM) (here in the notebook that incorporates stem-family size, word length, and distance to the stem-suffix boundary as predictors for the prediction measure.

Word length Both prediction and learning are dependent on word length [1]: short words are comparatively more difficult to predict, while they are easily learned. The dataset provided for this assignment did not include the training epochs, but we can observe the correlation between prediction and word length in Figure 7.

Learning effect The dataset provided for this assignment did not include the training epochs, but again in

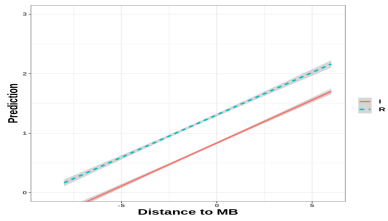

Figure 2:

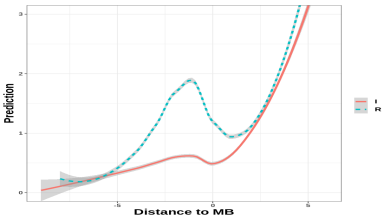

Figure 3:

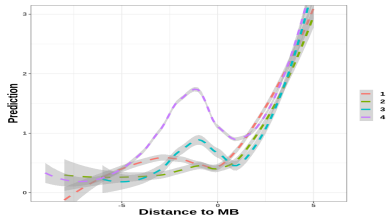

Figure 4:

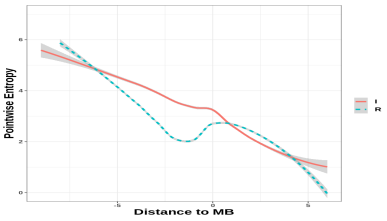

Figure 5:

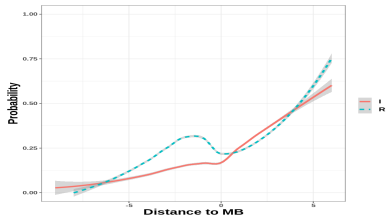

Figure 6:

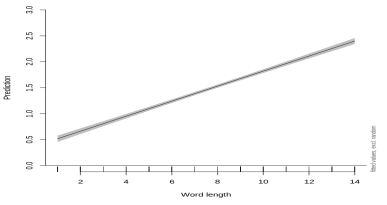

Figure 7:

Figure 2: Linear interaction of TSOM prediction rates in word recoding by letter position, with the categorical regular (blue) vs. irreguar(red) classification of paradigms;
Figure 3: Non-linear interaction obtained by using R ggplot2 package; smoothing is performed using the LOESS method; shaded areas represent the standard error
Figure 4: Non linear interaction, again using LOESS and grouping paradigms in (ir)regularity ranges
Figure 5: Non linear interaction plotting pointwise entropy against distance to MB
Figure 6: Non linear interaction plotting Probability against distance to MB
Figure 7: GAM model plot showing the relation between prediction and word length

[1] chain specialization through epochs is observed, following different paces cross-linguistically.

6. Conclusions

The structure of a language is shaped by the speaker's convenience, aiming to strike a balance between the ease of learning and the simplicity of using the system. Morphologically, this equilibrium is achieved through gradual regularity. On one hand, maximal contrast in inflectional systems facilitates word processing, allowing for the prompt disambiguation of verb forms, emphasizing early lexical discrimination albeit at the cost of higher frequency required for settling. On the other hand, word learning is facilitated in a system that maximizes generalizability, where forms share stems and follow regular inflections. Thus, the inflectional system must navigate between these two language-independent requirements for learnability, a perspective that aligns with psycholinguistic principles of word coactivation and competition.

References

[1] C. Marzi, M. Ferro, V. Pirrelli, A processing-oriented investigation of inflectional complexity, Frontiers in Communication 4 (2019). URL: https://www.frontiersin.org/articles/10.3389/fcomm.2019.00048. doi:10.3389/fcomm.2019.00048.

[2] M. Ferro, C. Marzi, V. Pirrelli, A self-organizing model of word storage and processing: Implications for morphology learning, Lingue e Linguaggio 2 (2011).

[3] T. Kohonen, The self-organizing map, Proceedings of the IEEE 78 (1990) 1464–1480. doi:10.1109/5.58325.

[4] C. Marzi, M. Ferro, V. Pirrelli, Morphological structure through lexical parsability, Lingue e linguaggio, Rivista semestrale (2014) 263–0. URL: https://www.rivisteweb.it/doi/10.1418/78410. doi:10.1418/78410.

[5] S. Rorberi, Un modello computazionale dell'interazione tra regolarità e struttura morfologica: il caso della flessione verbale del russo come l1 e l2, 2019.

[6] C. Marzi, Modelling the interaction of regularity and morphological structure: the case of russian verb inflection, Lingue e linguaggio, Rivista semestrale (2020) 131–156. URL: https://www.rivisteweb.it/doi/10.1418/97534. doi:10.1418/97534.

[7] S. Pinker, M. T. Ullman, The past and future of the past tense, Trends in Cognitive Sciences 6 (2002) 456–463. URL: https://www.sciencedirect.com/science/article/pii/S1364661302019903. doi:https://doi.org/10.1016/S1364-6613(02)01990-3.

[8] F. Basso, Un approccio computazionale alla complessità della derivazione: il caso dei derivati italiani in -zione, -tore, -ico, 2021.