# DISCRIMINATIVE WORD LEARNING IS SENSITIVE TO INFLECTIONAL ENTROPY

MARCELLO FERRO     CLAUDIA MARZI     VITO PIRRELLI

ABSTRACT: Psycholinguistic evidence based on inflectional and derivational word families has emphasised the combined role of Paradigm Entropy and Inflectional Entropy in human word processing. Although the way frequency distributions affect behavioural evidence is clear in broad outline, we still miss a clear algorithmic model of how such a complex interaction takes place and why. The main challenge is to understand how the local interaction of learning and processing principles in morphology can result in global effects that require knowledge of the overall distribution of stems and affixes in word families. We show that principles of discriminative learning can shed light on this issue. We simulate learning of verb inflection with a discriminative recurrent network of specialised processing units, whose level of temporal connectivity reflects the frequency distribution of input symbols in context. We analyse the temporal dynamic with which connection weights are adjusted during discriminative learning, to show that self-organised connections are optimally functional to word processing when the distribution of inflected forms in a paradigm (Paradigm Entropy) and the distribution of their inflectional affixes across paradigms (Inflectional Entropy) diverge minimally.

KEYWORDS: discriminative learning, word processing, recurrent neural networks, relative entropy.

## 1. INTRODUCTION[1]

Families of morphologically-related words, be they word paradigms (inflected forms of the same lemma), inflectional series (identically-inflected forms of different lemmas), derivational families (morphologically-complex words sharing the same root) or derivational series (morphologically-complex words

[1] Authors are alphabetically ordered. Marcello Ferro developed the TSOM software and ran experiment 1 and experiment 2 on Italian and Greek; Claudia Marzi ran experiment 2 on German and Spanish, conducted data analysis and statistical modelling of the results; Vito Pirrelli framed the theoretical and mathematical background. Implications and concluding remarks were jointly discussed. Claudia Marzi and Vito Pirrelli critically revised the paper.

sharing the same derivational affix), have received increasing attention over the last 25 years. Considerable emphasis has been laid on the role of paradigmatic relations as principles of non-linear organisation of word forms in the speaker's mental lexicon, facilitating their access, retention and use (Baayen et al. 1997; Orsolini & Marslen-Wilson 1997; Bybee & Slobin 1982; Bybee & Moder 1983, among others).

A large body of cognitive literature on similarity-based principles of word co-activation and competition has focused on effects of family size and frequency of neighbouring words on a variety of word processing tasks (Gathercole et al. 1997; Luce 1986; Luce & Pisoni 1998; Pitt & McQueen 1998; Vitevitch et al. 1997; Vitevitch & Luce 1998), to highlight an interesting general pattern. Large neighbour families tend to have facilitative effects on tasks like spoken word production and visual word recognition, but facilitation strongly interacts with frequency distributions of family members. Given a written word to be recognised, low-frequency neighbours facilitate both visual recognition and production, but high-frequency neighbours exert an inhibitory effect on the same tasks.

More recently, the study of word families prompted a growing interest in information-theoretic measures of their structure and organisation (e.g. Ackerman, Blevins & Malouf 2009). The interactive role of intra-paradigmatic and inter-paradigmatic word distributions has been systemically investigated to account for their differential effects on visual lexical recognition of both inflected (Milin et al. 2009a, 2009b) and derived words (see Kuperman et al. 2010; Bertram et al. 2000; Schreuder et al. 2003, among others). In particular, Milin and colleagues (2009a, 2009b) focus on the divergence between the distribution of inflectional endings within a single paradigm (measured as the entropy of the distribution of paradigmatically-related forms, or Paradigm Entropy), and the distribution of the same endings within their inflectional class (or Inflectional Entropy). They observe that both paradigm entropy and inflectional entropy facilitate visual lexical recognition: the more uniform the frequency distribution of word forms is in either family, the easier they are to process. However, if the two distributions differ, a conflict arises, resulting in slower word recognition. The difference between paradigm entropy and inflectional entropy is expressed in terms of Relative Entropy through the Kullback-Leibler Divergence (or *$D_{KL}$*, Kullback 1987), as follows:

1) $$D_{KL}\big((e\,|s)||p(e)\big) = \sum_e p(e \mid s) log \frac{p(e|s)}{p(e)}$$

where $p(e|s)$ is the probability of having a specific inflected form (an ending $e$) given a stem $s$, and $p(e)$ is the probability of finding $e$ with any $s$. For each paradigm, the larger 1) is, the more difficult is, on average, the visual recognition of members of that paradigm. Similar results are reported by Kuperman and colleagues (2010) on reading times for Dutch derived words, and are interpreted as reflecting an information imbalance between the family of the base word (e.g. *plaats* in *plaatsing*) and the family of the suffix (*-ing*).

All these effects are clear in broad outline. They point to a deeply rooted interaction between word distributions and word competition in the mental lexicon, where inflected forms are concurrently memorised, and synchronously accessed to compete for primacy in processing. Nonetheless, we still miss an algorithmic characterisation of the ways local storage and local processing functions make the human word processor exquisitely sensitive to global frequency effects. Computational models appear to be an ideal tool of the trade to investigate these and other related issues. They can provide a detailed, data-driven account of the spontaneous emergence of sensitivity to frequency effects from patterns of language usage. Experiments conducted by implementing and running computer simulations of a specific language task can be used to understand more of processing behaviour by testing the principled, cognitively-grounded mechanisms that are assumed to be the cause of this behaviour, but are inaccessible to classical psycholinguistic experiments.

In what follows, we intend to offer and empirically validate a full-fledged, connectionist model of these effects, based on principles of discriminative learning (Rescorla & Wagner 1972), implemented through recurrent self-organising neural networks (Ferro et al. 2011; Marzi et al. 2014; Pirrelli et al. 2015). Section 2 introduces some fundamental information-theoretic equations. Section 3 illustrates the principles governing discriminative learning and their neural network implementation. Section 4 describes experimental protocols, materials and results. The theoretical implications of our model are discussed in the concluding section.

## 2. PARDIGMS AND ENTROPY

By way of illustration, we consider two artificial mini-paradigms, obtained by combining two stems (‘A’ and ‘B’) with two endings (‘X’ and ‘Y’). Table 1 shows paradigm and inflectional entropic scores of the artificial inflectional system. Members of the mini-paradigms are distributed according to the frequencies in Table 1.1. The probability of encountering any inflected form is expressed by the joint probability $p(s_k, e_h)$ of finding the stem $s_k$ followed by the inflectional ending $e_h$, and is calculated as the ratio $f(s_k, e_h)/$

$\sum_{i,j} f(s_i, e_j)$ (Table 1.2, grey cells). Accordingly, we can express the probability of selecting one particular inflected form from its own paradigm as $p(e_h|s_k)$, which is the conditional probability of finding one particular ending $e_h$ given the stem $s_k$, or $p(s_k, e_h)/p(s_k)$ (Table 1.3). Note that $p(e_h|s_k)$ equals $p(e_h)$ when knowledge of $s_k$ does not reduce the uncertainty about $e_h$, that is, if the two events are independent and $p(s_k, e_h) = p(s_k) \cdot p(e_h)$. Incidentally, this is the case of our distribution in Table 1. Similarly, $p(s_k|e_h)$ is the conditional probability of the stem $s_k$ given $e_h$, i.e. the probability that $s_k$ is found with $e_h$, when we restrict ourselves to the words ending in $e_h$ only (Table 1.4). The distribution of $p(s_k|e_h)$ gives information about how forms are distributed within a paradigm cell. Once more, $p(s_k|e_h)$ equals $p(s_k)$ when $s_k$ and $e_h$ are distributed independently.

1)

| | X | Y | $f(s)$ |
|---|---|---|---|
| A | 56 | 14 | 70 |
| B | 24 | 6 | 30 |
| $f(e)$ | 80 | 20 | *100* |

2)

| | X | Y | $p(s)$ |
|---|---|---|---|
| A | 0.56 | 0.14 | 0.7 |
| B | 0.24 | 0.06 | 0.3 |
| $p(e)$ | 0.8 | 0.2 | *1* |

3)

| | $p(X \mid s)$ | $p(Y \mid s)$ | $H(e \mid s_i)$ |
|---|---|---|---|
| A | 0.8 | 0.2 | 0.72 |
| B | 0.8 | 0.2 | 0.72 |
| $H(e \mid s)$ | | | *0.72* |

4)

| | X | Y | $H(s \mid e)$ |
|---|---|---|---|
| $p(A \mid e)$ | 0.7 | 0.7 | |
| $p(B \mid e)$ | 0.3 | 0.3 | |
| $H(s \mid e_i)$ | 0.88 | 0.88 | *0.88* |

Table 1. Paradigm-based word distributions illustrated with 2 mini-paradigms of 2 forms each. Frequency distributions (1) are transformed into word probabilities (2), paradigm probabilities (3) and inflectional class probabilities (4).

For all these probability distributions, we can calculate their respective entropy, and measure how uniform they are. We start from Table 1.2, with the stem entropy $H(s)$, defined as:

2) $H(s) = -\sum_j p(s_j) log\, p(s_j) = 0.88$

Similarly, the inflectional entropy $H(e)$ is given by:

3) $H(e) = -\sum_j p(e_j) log\, p(e_j) = 0.72$

The paradigm entropy $H(e|s_k)$ is calculated in Table 1.3 according to equation 4):

4) $H(e|s_k) = -\sum_j p(e_j|\, s_k) log\, p(e_j|s_k)$

By averaging the sum of the paradigm entropies of our two mini-paradigms (weighted by their probability $p(s)$), we get (Table 1.3):

5) $H(e|s) = \sum_i p(s_i)\, H(e|s_i) = 0.72$

which is equal to $H(e)$ in equation 3), due to the distributional independence between stems and endings. The cell entropy $H(s|e_k)$ (or entropy of identically-inflected words) measures the distribution of all inflected forms ending in $e_k$ as follows:

6) $H(s|e_k) = -\sum_i p(s_i|\, e_k) log\, p(s_i|e_k)$

Its averaged sum over all endings in an inflection class, $H(s|e)$, is calculated in Table 1.4 according to equation 7) below:

7) $H(s|e) = \sum_i p(e_i)\, H(s|e_i) = 0.88$

Again, due to the independence condition, $H(s|e)$ equals $H(s)$ in equation 2). Finally, $H(s,e)$ is the entropy of the $p\left(s_i, e_j\right)$ distribution of full inflected forms, and is calculated thus:

8) $H(s,e) = -\sum_{i,j} p\left(s_i, e_j\right) \log p\left(s_i, e_j\right) = 1.6$

Figure 1 is a diagrammatic representation of the relations between $H(s,e)$, $H(s|e)$, $H(e|s)$ and $I(s,e)$. $I(s,e)$ (known as Mutual Information) is a measure of the mutual dependence between stems and endings, defined as the divergence of the distribution $p(s,e)$ from the independence hypothesis $p(s,e) = p(s) \cdot p(e)$ (Manning & Schütze 1999):

9) $I(s,e) = \sum_{i,j} p(s_i, e_j) log \frac{p(s_i, e_j)}{p(s_i)p(e_j)}$

Using the definition of conditional probability, we can replace $p(s,e)$ in equation 9) with $p(s) \cdot p(e|s)$, to obtain:

10) $I(s,e) = \sum_i p(s_i) \sum_j p(e_j \mid s_i) log \frac{p(e_j|s_i)}{p(e_j)}$

where $\sum_j p(e_j \mid s_i) log \frac{p(e_j|s_i)}{p(e_j)}$ is the relative entropy of equation 1) above. $I(s,e)$ can thus be interpreted as the averaged divergence between the distribution of inflected forms within their paradigms ($p(e|s)$) and the distribution of their inflectional endings across all paradigms ($p(e)$). In our example, the two distributions are identical, hence $I(s,e)$ is empty. This means that knowledge of the stems gives no information about the distribution of the endings. Equivalently, we can say the stems exhibit no particular selection preference for a specific subset of endings.

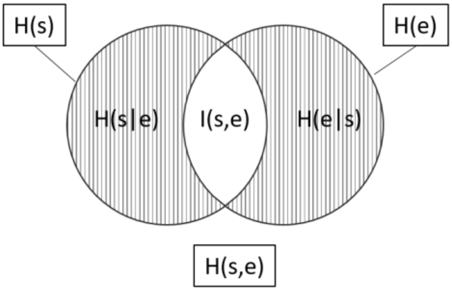

FIGURE 1. A DIAGRAMMATIC REPRESENTATION OF THE RELATIONS BETWEEN H(S,E), H(S|E), H(E|S) AND I(S,E), IN TERMS OF SET INTERSECTION AND SUBTRACTION RELATIONS. I(S,E) IS A MEASURE OF THE DIVERGENCE BETWEEN THE DISTRIBUTION OF THE INFLECTED FORMS IN EACH PARADIGM AND THE DISTRIBUTION OF THE ENDINGS IN THE CORRESPONDING INFLECTIONAL CLASS, AVERAGED OVER PARADIGM PROBABILITIES.

Finally, if we replace $p(s,e)$ with $p(e) \cdot p(s|e)$ in equation 9), we get:

11) $I(s,e) = \sum_i p(e_i) \sum_j p(s_j \mid e_i) log \frac{p(s_j|e_i)}{p(s_j)}$

where $\sum_j p(s_j \mid e_i) log \frac{p(s_j|e_i)}{p(s_j)}$ is the D<sub>KL</sub> divergence between the probability distribution $p(s)$ of the stems (rightmost column in Table 1.2) and the distribution of the same stems in identically-inflected forms (shaded columns in Table 1.4). As shown in Figure 1, $H(s|e) \leq H(s)$. Once more, $H(s|e) = H(s)$ when $\mathrm{D}_{KL} = 0$, i.e. when endings predict nothing about the distribution of stems.

To sum up, $I(s,e)$ can be interpreted in two symmetrical ways. Firstly, It is the distance between the paradigm entropy $H(e|s)$ and the inflectional entropy $H(e)$. Secondly, it measures the distance between the cell entropy $H(s|e)$ and the stem entropy $H(s)$. In fact, $I(s,e)$ quantifies the information that *s* and *e* share, i.e. how much knowing either variable reduces uncertainty about the other. Thus, it works in both directions. In particular, if $p(e_j \mid s_i) < p(e_j)$, i.e. if an inflected form occurs in a paradigm less frequently than one would expect considering the frequency of its inflectional ending, then $p(s_i \mid e_j) < p(s_i)$, i.e. its stem will be under-represented in the corresponding set of identically-inflected forms. As we shall see in more detail in section 3, this has important repercussions on the learning behaviour of a recurrent neural network where forms are concurrently stored and accessed as a function of $H(s|e)$.

## 3. DISCRIMINATIVE WORD LEARNING

From a discriminative perspective, learning proceeds by discriminating between multiple cues that are constantly in competition for their predictive value for a given outcome (Ramscar & Yarlett 2007; Baayen et al. 2011; Blevins 2016). We conjecture that the sensitivity of human word processing to effects of paradigm relative entropy reflects the dynamic interaction between concurrently stored items, due to the superpositional nature of their stored representations and the dual function of these representations as processing units.

Work in discriminative word learning has primarily focused on form-meaning relationships based on highly distributed amorphous representations. A recurrent network variant of discriminative learning was recently used with one-level self-organising grids of processing nodes known as Temporal Self-Organising Maps (TSOMs, Ferro et al. 2011; Marzi et al. 2014; Pirrelli et al. 2015). TSOMs memorise time-series of symbols as chains of specialised processing nodes, selectively firing when specific symbols are input in specific temporal contexts. TSOMs consist of a bank of input nodes (where input stimuli are encoded), and a bank of processing nodes (the map proper), connected to input nodes through input connections, and to processing nodes (including themselves) through re-entrant temporal connections with one-time delay (Figure 2).

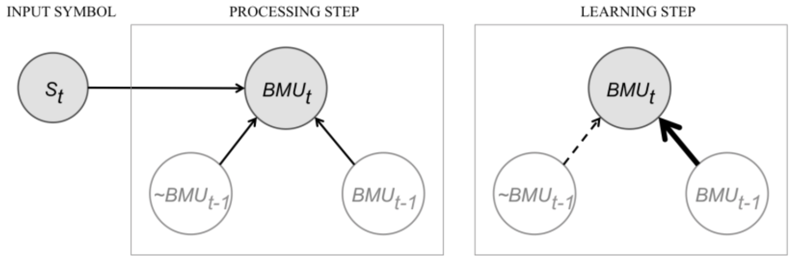

Figure 2. Activation (left) and learning (right) steps in a TSOM. 'BMU' and '~BMU' represent, respectively, a best matching unit and any other node that is not a best matching unit. Subscripts index time ticks.

At each time tick *t*, activation flows from the input layer to the map nodes through one-way input connections (Figure 2, left panel). Re-entrant temporal connections update each map node with the state of activation of all nodes at the previous time tick (*t-1*). Like with classical Recurrent Neural Networks (Elman 2009), a word is input to a TSOM one symbol *S* at a time. Activation spreads through both input and temporal connections to yield an overall state of node activation, or Map Activation Pattern for *S* at time *t*: $MAP_t$ (*S*). The node with the top-most activation level in $MAP_t$(*S*) is called the Best Matching

Unit for *S* at time *t*, or $BMU_t(S)$. A time series of sequentially activated *BMUs* will hereafter be referred to as a *BMU* chain.

Weights on temporal connections encode how strongly the current $BMU_t$ is predicted by $BMU_{t-1}$, ranging continuously from 0 to 1. Temporal connection weights are trained on input data according to the following principles of correlative learning, strongly reminiscent of Rescorla & Wagner (1972) discriminative equations (Figure 2, right panel). Namely, when the bigram 'AX' is input, a TSOM goes through two learning steps:

(i) the temporal connection between $BMU_{t-1}(A)$ and $BMU_t(X)$ (upwards thick arrow in Figure 2, right) is strengthened (entrenchment);

(ii) all other temporal connections to $BMU_t(X)$ (dashed arrow in Figure 2, right) are weakened (competition).

It is useful to consider the effects of the learning steps in some detail. Owing to step (i), connection strength increases as a function of how often the connection is traversed *leaving* a particular node. Interaction with step (ii), however, makes strengthening conditional on how often a connection is traversed to *arrive* at a specific node. This is illustrated in Figure 3, which provides a graph-like representation of *BMU* chains trained on the probability distribution of the mini-paradigms in Table 1 above. In the left panel, transition probabilities are conditional on stems: for example, $p(X|A) = 0.8$ says how much of the probability mass of 'A' is channelled through the 'X' connection. In the right panel, probabilities are conditional on endings: $p(A|X) = 0.7$ says how much of the probability of 'X' comes from 'A'. This is useful to understand how competition effects are modulated by frequency distributions.

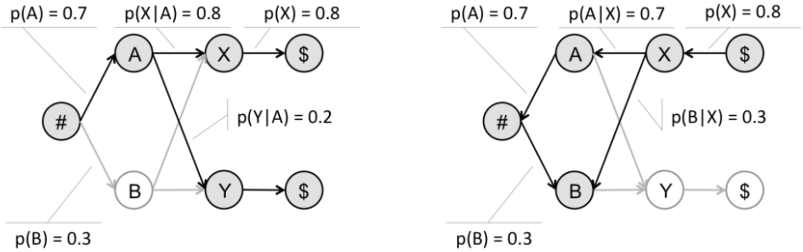

FIGURE 3. A GRAPH-LIKE REPRESENTATION OF BMU CHAINS TRAINED ON TWO MINI-PARADIGMS, ACCORDING TO THE FORWARD (LEFT) AND BACKWARD (RIGHT) PROBABILITY DISTRIBUTIONS OF TABLE 1. '#' AND '$' ARE, RESPECTIVELY, THE START-OF-WORD AND END-OF-WORD SYMBOLS.

According to step (ii), the strength of the A-X connection is conditional on the occurrence of the 'X' event, and is modelled by the backward probability distribution $p(A|X)$. This means that the strength of A-X is a function of how often 'A' occurs with 'X' compared with how often 'X' follows any other stem $s$: namely $p(A,X)/p(X)$. The entropy $H(s|X)$ thus measures a memory effect. It says how many different stems the node 'X' can combine with: the larger the number, the higher $H(s|X)$. As a result of this dynamic, a map node with a high backward entropy tends to be less predictive than a node with a low backward entropy, as the former must keep memory of many preceding contexts and thus expects more possible forward continuations of these contexts.[2]

In a combinatorial system like verb inflection, being predictive is also a function of regularity. In regular paradigms, stems combine with endings more systematically than in irregular paradigms. For this reason, they turn out to be more uncertain to process at the stem-ending boundary than irregular stems are, due to the larger number of endings they are followed by. Irregular stems are more predictive, as they typically select a specific subset of endings only. Hence, the forward entropy $H(e|s)$ of a regular stem is higher than the forward entropy of a stem allomorph in an irregular paradigm.

Backward and forward entropies are not the only factors affecting word processing in TSOMs. Although nodes with higher forward entropies are less predictive than nodes with lower forward entropies are, the distribution of endings within a paradigm is predicted more accurately when it is close to the distribution of endings across the entire inflectional system. To understand why, it is useful to remind that, owing to learning step (ii), the strength of each stem-ending connection is competitively affected by the probability mass of all other stems selecting the same ending. The probability distribution $p(s)$ of all stems is, in fact, a weighted centroid of the probability distributions $p(s|e)$ of the same stems in their paradigm cells. The closer $p(s|e)$ is to $p(s)$, the more evenly an inflected form in a given paradigm is competing with identically inflected forms in other paradigms. The cell entropy $H(s|e)$ measures this level of competition, with higher values corresponding to a more balanced competition. In particular, we know from Figure 1 that:

12) $$H(s|e) = H(s) - I(s,e)$$

where, for $H(s)$ being held constant, $H(s|e)$ goes up as $I(s,e)$ approaches 0. This means that the most balanced competition between *BMU* chains responding to identically inflected words is obtained when stems and endings are independently distributed (see section 2). As we change the probability distribution of Table 1, while keeping the marginal probabilities $p(s)$ and $p(e)$ fixed,

[2] TSOMs are biased towards learning the most discriminative such chains, i.e. those chains where each node is preceded by the smallest possible number of equiprobable nodes, given the resources available. This is achieved through context-sensitive specialisation. Memory resources allowing, symbols that are preceded by different contexts tend to be processed by different, specialised *BMUs*.

we shift away from an optimally balanced competition between inflected forms, towards a suboptimal distribution where some forms compete more or less strongly than one would expect considering their stem probability.

Finally, due to the symmetry of the relationships depicted in Figure 1, also the following equation is true:

13) $H(e|s) = H(e) - I(s, e)$

By varying $I(s, e)$ we are in fact also varying, by the same amount, the distance between the distribution of the endings in their inflectional class and the distribution of the same endings within each verb paradigm.

## 4. MATERIALS AND METHODS

To assess the role of relative entropy in the processing of paradigm-based inflectional systems, we ran two experiments. In the first experiment, a TSOM was trained on a set of artificial mini-paradigms, whose frequency distribution was varied to control for $H(e)$ and $H(e|s)$. For each different distribution, we repeated a complete training session 100 times, and assessed trained TSOMs both structurally, i.e. in terms of levels of temporal connectivity, and functionally, i.e. by looking at their processing behaviour. The experiment aimed to highlight main trends, and disentangle hierarchical factor interactions.

In the second experiment, we looked for similar trends in the inflectional systems of four different languages: German, Italian, Modern Greek and Spanish. A homogenous sample of 50 sub-paradigms was selected among the most highly frequent paradigms in each language. Frequency distributions were made vary across two different training protocols. To control for random variability, the same protocol was repeated 5 times.

Results were analysed using Generalised Additive Models (GAM), with probability and entropy distributions of training data as predictors.

### *4.1 Experiment 1*

Starting with an artificial set of three mini-paradigms, we created 729 different training regimes with all possible combinations of three frequency bins (10, 100, 1000) in a contingency table like Table 2, which illustrates one such possible combination. The experiment consisted in training a TSOM on each frequency bin combination.

| | X | Y | *f(s)* |
|---|---|---|---|
| A | 10 | 100 | 110 |
| B | 100 | 100 | 200 |
| C | 1000 | 10 | 1020 |
| *f(e)* | 1110 | 210 | *1320* |

TABLE 2. ONE COMBINATION OF THREE FREQUENCY BINS FOR THREE MINI-PARADIGMS.

Figure 4 shows how the general trends discussed in section 3 affect the development of TSOM connectivity across different training regimes. First, an even distribution of inflectional endings in the input brings about a balanced apportioning of connection weights at the stem-ending boundary, which grow in strength as the inflection entropy *H*(*e*) gets higher (left panel), owing to the learning step (ii).

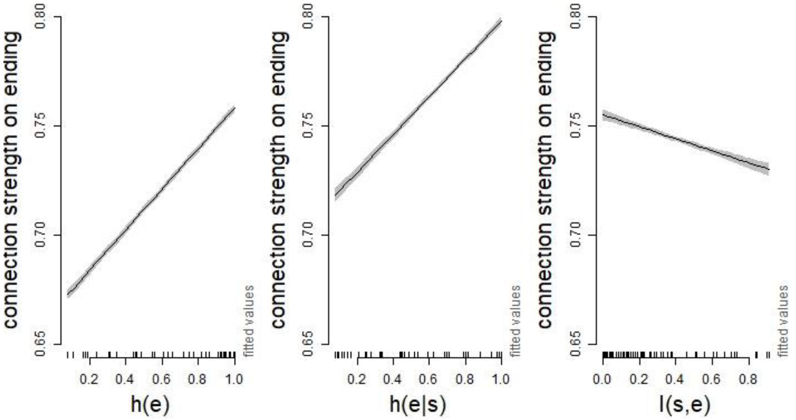

FIGURE 4. GAM PREDICTING CONNECTION STRENGTH AT THE MORPHEME BOUNDARY. FIXED EFFECTS ARE PLOTTED SEPARATELY AS *H*(*E*) (LEFT PANEL), *H*(*E*|*S*) (CENTRAL PANEL), AND *I*(*S*,*E*) (RIGHT PANEL).

The same holds for the distribution of stems across paradigm cells (central panel). For higher values of *H*(*e*|*s*), a TSOM develops stronger connections. Once more, we observe a family size effect here. When paradigm members are evenly distributed, their corresponding node chains are better allocated and, processing resources allowing, also more discriminative. A more skewed distribution of the same families has a statistically significant inhibitory effect on connection strengths (p-value <.001). This is confirmed by the negative slope for growing values of relative entropy averaged across all paradigms in our training data, as expressed by the Mutual Information *I*(*s*,*e*).

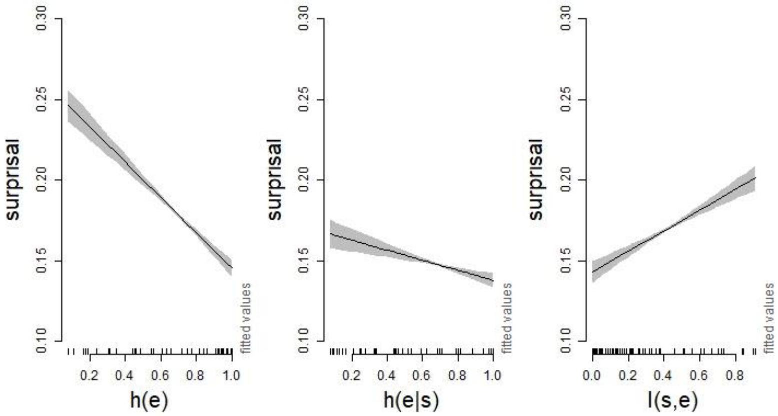

FIGURE 5. GAM PREDICTING PROCESSING SURPRISAL. FIXED EFFECTS ARE PLOTTED SEPARATELY AS $H(E)$ (LEFT PANEL), $H(E|S)$ (CENTRAL PANEL), AND $I(S,E)$ (RIGHT PANEL).

Figure 5 illustrates the neat effects of this structural dynamic on processing the mini paradigms. Here, predictors are related to levels of processing "surprisal" (Levy 2008), a robust, information-theoretic measure of how unexpected an input symbol is on the basis of its immediately preceding context.[3] Surprisal is demonstrably correlated with differential, local processing complexity: the higher its value, the more difficult for the map to process a word at a particular position in the input string. As the distribution of input endings get increasingly uniform across training regimes (left panel), the stem-ending transition is likely to be associated with more entrenched node chains. In a similar vein, higher values of paradigm entropy $H(e|s)$ make processing less uncertain at the stem-ending boundary (central panel). Once more, the result is consistent with the idea that a low-entropy paradigm is governed by the distribution of few of its members only, which take most of the processing resources that the TSOM allocates for the whole paradigm. On average, this increases the surprisal in processing paradigm members with low $p(e|s)$. In addition, a TSOM finds it increasingly more difficult to process paradigms that are more off-centred with respect to the general distributional tendency $H(e)$ of their inflectional class. This is shown by the right panel of Figure 5, where $I(s,e)$ measures the average distance of paradigm distributions from $H(e)$ (see section 2). Intuitively, this means that when the distribution of a paradigm diverges from the distribution of its inflectional class, its forms suffer from the competitive pressure of the majority of identically inflected words of other paradigms.

[3] In its basic form, surprisal is defined as the negative log-probability of the symbol $s_i$ given its left context, or $-logp(s_i|s_{1,\dots,i-1})$. In a TSOM, $p(s_i|s_{1,\dots,i-1})$ is approximated by the temporal level of activation of $BMU(s_i)$ divided by the overall level of temporal activation of the map at the same time tick.

### *4.2 Experiment 2*

We used four training sets of German, Italian, Modern Greek and Spanish verb forms. For each language, the set consists of the 50 top-frequency verb subparadigms of 15 inflected forms, selected from a common pool of paradigm cells including present (*n*=6) and past (*n* =6) indicative tenses, the infinitive (*n* =1), past participle and gerund/present participle (*n* =2). A TSOM was trained on each set under two training regimes (Marzi et al. 2016, 2018). In the first regime, forms were presented with a uniform frequency distribution, inputting each item 5 times. In the second regime, the same set of forms was presented with realistic frequency distributions, sampled from reference corpora. The two training regimes were simulated 5 times. Frequency distributions and simulation results were collected and normalised for each language separately.

Our training data included both regular and irregular inflection, in different proportions for the four languages, in the two training conditions. To control for allomorphy, paradigm distributions were based on counting specific combinations of stem and affix alternants. In irregular paradigms, this way of counting resulted in much finer-grained morphological families than traditional paradigms and inflectional classes. In line with information-theoretic approaches to paradigm-based inference, our distributions reflect the way (mostly irregular) paradigms are partitioned into smaller classes of mutually implied inflected forms, indexed by a formally unique stem allomorph (Pirrelli 2000; Stump 2001). Under this view, irregular paradigms consist of a collection of stem partition classes (Pirrelli 2000), which are defined as families of inflected forms that share the same stem allomorph, and select a subset of inflectional endings. Regular paradigms, on the other hand, are families of inflected forms containing one such partition class only. To illustrate, in the Italian irregular paradigm VENIRE 'come', the diphthongised stem allomorph *vien-* (as opposed to default *ven-*) is only found in the second singular (*vieni*, 'you come') and third singular (*viene*, '(s)he comes') forms of the present indicative. In particular, in our training set, $p(\textit{vieni}|\text{VIEN-}) = 0.01$ and $p(\textit{viene}|\text{VIEN-}) = 0.99$. Hence, the corresponding paradigm entropy (0.09) is low, and its relative entropy (3.38) is twice as much as the average relative entropy of a regular paradigm (1.56). This way of operationalising stem partition classes allows us to capture gradient levels of irregularity in complex inflectional systems.

The plots in Figure 6 parallel those in Figure 4 for mini-paradigms.[4] As expected, increasing values of $p(e|s)$ strengthen the connections to nodes responding to inflectional endings (left panel).[5] This is a straightforward result of memory entrenchment, modelled by learning step (i). Like in Experiment 1, $H(e|s)$ has a statistically significant facilitative effect (p-value<.001) on the entrenchment of stem-ending connections (Figure 6, central panel). For comparable values of $p(e|s)$, inflected forms in high-entropy paradigms develop stronger connections at the stem-ending boundary. We can explain this as a paradigm regularity effect. In regular paradigms, stem partition classes contain more forms and more evenly distributed ones than they do in irregular paradigms. In our training set, paradigm entropy in regular paradigms is significant larger than in irregulars for all four languages (p-values< .001). This implies that forms in regular paradigms enter a more balanced competition than irregulars do, and this favours word acquisition and processing.

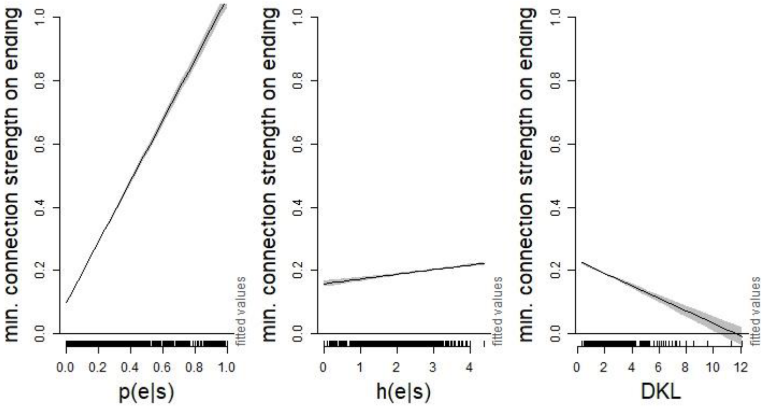

FIGURE 6. GAM PREDICTING THE MINIMUM VALUE OF CONNECTION STRENGTH ON ENDINGS. FIXED EFFECTS ARE PLOTTED SEPARATELY AS $P(E|S)$ (LEFT PANEL), $H(E|S)$ (CENTRAL PANEL), AND $D_{KL}$ (RIGHT PANEL).

The inhibitory effect of relative entropy ($D_{KL}$, right panel in Figure 6) is another, subtle consequence of family-based competition. Those paradigms whose distributions are closer to the distribution of the endings in the corresponding inflectional class are likely to develop more entrenched node chains.

[4] Unlike in Experiment 1, the data points of Experiment 2 do not represent the average behaviour of a TSOM for each distinct training regime, but the response of a TSOM to individual inflected forms for the four languages considered.

[5] Here, we select the node with the minimum incoming connection strength on incoming connection. Elsewhere (Marzi et al. 2018), we showed that this node typically marks the stem-ending boundary. Nonetheless, its position may occasionally vary, depending on the possible presence of linking elements between the stem and the inflectional ending proper (e.g. thematic vowels).

As shown in section 2, $D_{KL}$ measures the distributional dependence between stems and endings, and, ultimately, the extent to which each stem competes with all other stems in the inflectional class. It is interesting to note that this dependence correlates with inflectional regularity. In our training set, regular paradigms exhibit significantly lower levels of $D_{KL}$ than irregulars do (p-value < .001). In addition, high entropy paradigms show low $D_{KL}$.

Figure 7 plots the processing effects of these structural trends. Here, $p(e|s)$ shows a clear facilitative effect (left panel). Due to entrenchment, more likely inflected forms in their paradigms recruit more processing resources (i.e. more dedicated node chains), and this lowers processing surprisal. Likewise, the processing of higher entropy paradigms (central panel in Figure 7) is slightly but significantly facilitated (p-value <.001). Finally, relative entropy (right panel in Figure 7) has a rather clearer effect on processing complexity. When the distribution of the forms of a paradigm diverges from the distributional centroid of its inflectional class, processing surprisal significantly increases (p-value <.001).

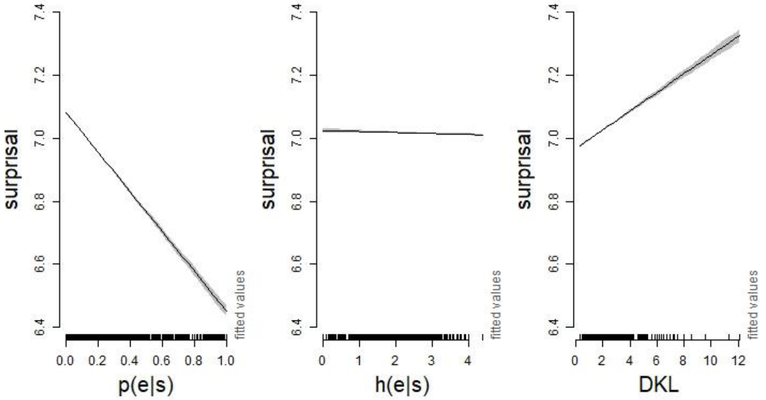

FIGURE 7. GAM PREDICTING PROCESSING SURPRISAL. FIXED EFFECTS ARE PLOTTED SEPARATELY AS $P(E|S)$ (LEFT PANEL), $H(E|S)$ (CENTRAL PANEL), AND $D_{KL}$ (RIGHT PANEL).

Notably, the same trends are confirmed by two distinct GAMs predicting connection strength and processing surprisal in the uniform training regime and in the corpus-based one. Clearly, differences in entropic scores are damped in the uniform regime, but they nonetheless reached statistical significance with all predictors. This shows that the effects are rooted in deep, structural differences in the data, amenable to the categorical distinction between regular and irregular paradigms. We will return to this point in the general discussion.

## 5. GENERAL DISCUSSION

Competition among multiple lexical cues for their discriminative value has recently been shown to be key to accounting for fundamental aspects of word processing (Baayen et al. 2011; Ramscar & Yarlett 2007; Ramscar et al. 2013; Milin et al. 2017). In the majority of discriminative approaches to language learning we are aware of, units defined on one level of representation are understood and modelled to cue units on a different level. For example, forms are cues to either lexical or morpho-syntactic content in both Baayen et al. (2011) and Ramscar & Yarlett (2007). In the present paper, we showed that the same pool of equations going back to Rescorla & Wagner (1972) can model the way simple word forms (with no lexical or morphological content) are concurrently memorised in a self-organising recurrent neural network (TSOM), and compete for primacy during processing through co-activation. A TSOM uses discriminative equations to develop maximally efficient *BMU* chains, within a grid of self-organised processing units (nodes) with one level of re-entrant temporal connections. Notably, unlike other discriminative learning models, *BMU* chains are defined on one representation level only.

In TSOMs, competition is an effect of the distributed nature of memory representations based on node superposition/correlation. Partially overlapping members of the same family activate *BMU* chains that share identical nodes. For example, inflected forms belonging to the same paradigm (e.g. *walking* and *walked*), or filling the same paradigm cell (e.g. *walking* and *speaking*), trigger node-sharing *BMU* chains. Entrenchment of shared nodes benefits from cumulative exposure to redundant input patterns, making the network sensitive to systematic sublexical structures in the input (e.g. the stem *walk-* in *walking* and *walked*, or the ending *-ing* in *walking* and *speaking*). Thus, larger word families favour entrenchment of shared substructures. Conversely, non-shared nodes in partially overlapping memory traces compete for time-locked activation primacy in processing, due to the discriminative learning bias governed by steps (i) and (ii). Thus, other things being equal, their temporal connections are modulated by their competition and, ultimately, by the entropy $H(e|s)$ of their distribution. Higher paradigm entropies favour a balanced allocation of memory resources; so stem-ending connections that are traversed with the same probability turn out to be stronger in words whose paradigms are more highly entropic. This is in line with evidence of human word processing (e.g. Lively et al. 1994; Luce 1986; Luce & Pisoni 1998).

The same dynamic provides an algorithmic account of evidence that human visual word processing is facilitated when the distribution of the set of inflected forms of a given stem diverges minimally from the distribution of the inflectional endings in the stem's inflectional class (i.e. for low values of paradigm relative entropy). It should be reminded from the discussion of session 3 that, when the distribution of endings is independent of the distribution of stems, the entropy of the paradigm cell $H(s|e)$ is maximum. As a result, *s-e* connections compete on a par, and get evenly strong during learning. In addition, maximisation of $H(s|e)$ means that also $H(e|s)$ is maximum (see section 2).[6] This provides the crucial explanatory link between results of our simulations and Milin and colleagues' evidence on visual word recognition. Words in stem families with low relative entropy ($D_{KL}$) are processed more easily than words in families with higher relative entropy because the former minimise competition by both identically inflected words of other paradigms (maximum $H(s|e)$) and differently inflected forms in the same paradigm (maximum $H(e|s)$).

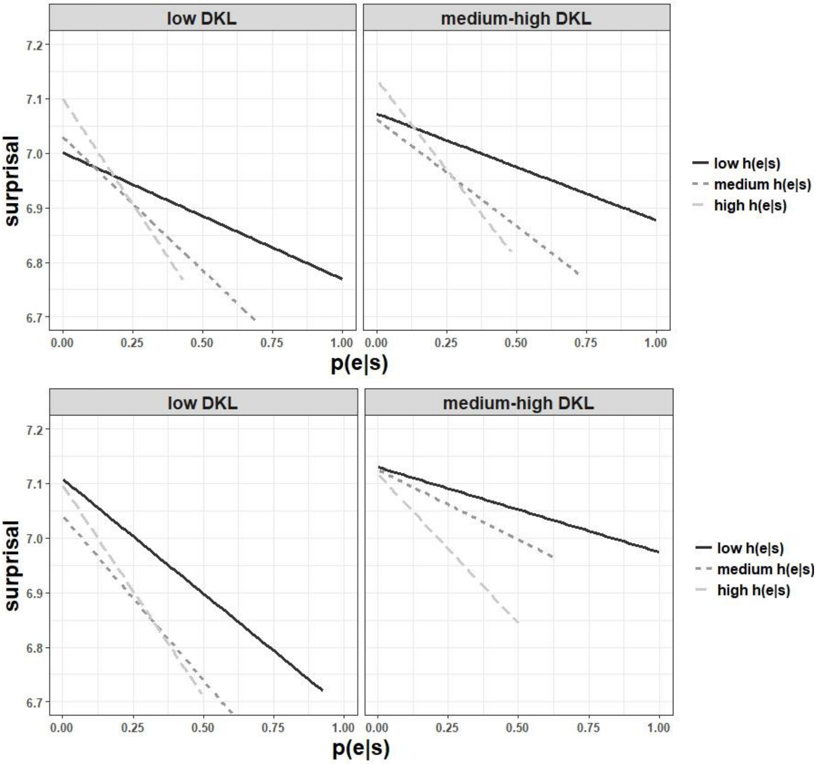

FIGURE 8. INTERACTION EFFECTS BETWEEN $P(E|S)$, $H(E|S)$ AND $DKL$ IN A GAM PREDICTING PROCESSING SURPRISAL FOR VERB FORMS IN IRREGULAR (TOP PANELS) AND REGULAR (BOTTOM PANELS) PARADIGMS IN CORPUS-BASED FREQUENCY DISTRIBUTIONS.

[6] Note, however, that the two values do not correlate, since endings can belong to different inflectional classes and can combine with partition classes of different size.

Figure 8 shows how processing surprisal varies in irregular (top panels) and regular (bottom panels) paradigms by levels of paradigm entropy ($H(e|s)$) and relative entropy ($D_{KL}$). In all four panels there is a tendency for surprisal to decrease for increasing levels of paradigm entropy. On average, surprisal is significantly higher for forms in regular paradigms than in irregular ones (p-value <.001). Nonetheless, the facilitative effect of low relative entropy is differently modulated by inflectional regularity: forms in regular paradigms benefit more from decreasing relative entropy than irregular forms do.[7]

In line with this evidence, inflectional regularity can be described as a tendency of word paradigms to approximate the central distribution of their inflectional classes. Regular paradigms are significantly closer to their inflectional centroid, and this gives them a processing advantage. Conversely, in irregular paradigms, inflectional endings are strongly selected by specific stems in small partition classes, and their paradigmatic distribution typically diverges from the central distribution of the whole verb system. They thus receive comparatively little global support from other paradigms, and their level of entrenchment is mainly governed by token frequencies, as opposed to family frequency effects.

To sum up, our evidence establishes a connection between the paradigm relative entropy effect and the family size effect (de Jong et al. 2000; Mulder et al. 2014), suggesting that the two are the by-product of the same underlying dynamic. In addition, our data highlights an interplay between processing-oriented effects and the categorical distinction between regular and irregular inflection. Being regular means being part of large word families, and this facilitates the entrenchment of overlapping *BMU* chains through discriminative learning. For instance, regularity prompts the development of dedicated chains for the invariant stem of a regular verb paradigm, or the shared ending of a paradigm cell. In addition, being regular also means being processed with more uncertainty at the stem-ending boundary, because a regular stem typically combines with all endings of its inflectional class. Nonetheless, we observed that if the paradigm distribution diverges minimally from the distribution of its endings, processing surprisal is reduced. This is due to the operation of discriminative learning, whereby a TSOM develops some general expectations about the central distribution of inflectional classes. These expectations facilitate the processing of regularly inflected forms, making it up for their combinatorial behaviour. As expected, the relative entropy effect is only residual in irregular verb paradigms, where stem allomorphy reduces uncertainty at the stem-ending boundary, skewing stem-ending distributions away from their central tendencies.

[7] The reducing effect on surprisal of low relative entropy is larger for regulars (slope coefficient -0.42, p-value <.001) than for irregulars (-0.23, p-value <.001).

In the end, both strategies are effective in reducing word processing uncertainty. One targets the more idiosyncratic, high frequency portion of a language conjugation: the set of irregularly inflected verbs. The other one deals with the long Zipfian tail of regulars, which are formed in a combinatorial way. In the former case, local frequency effects, based on local allomorphies, matter most. In the latter case, global distributional effects appear to carry more weight. It is remarkable that both strategies do not call for independent processing modules, but can follow from the local interaction of a single pool of discriminative learning principles.

## REFERENCES

Ackerman, F., Blevins, J.P., & Malouf, R. (2009). Parts and wholes: Implicative patterns in inflectional paradigms. In J.P. Blevins & J. Blevins (eds.) *Analogy in grammar: Form and acquisition*, 54-82. Oxford: Oxford University Press.

Baayen, R.H., P. Milin, D.F. Đurđević, P. Hendrix & M. Marelli (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological review* 118(3). 438.

Baayen, R.H., T. Dijkstra & R. Schreuder (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language* 37(1). 94-117.

Blevins, J.P. (2016). Word and paradigm morphology. Oxford University Press.

Bybee, J.L. & C.L. Moder (1983). Morphological classes as natural categories. *Language* 59(2). 251-270.

Bybee, J.L. & D.I. Slobin (1982). Rules and Schemas in the Development and Use of the English past tense. *Language* 58(2). 265-289.

de Jong, N.H., R. Schreuder & R.H. Baayen (2000). The morphological family size effect and morphology. *Language and cognitive processes* 15(4/5). 329-365.

Ellis, N.C. (2002). Frequency effects in language processing. *Studies in second language acquisition* 24. 143-188.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science* 33. 547-582.

Ferro M., C. Marzi & V. Pirrelli (2011). A Self-Organizing Model of Word Storage and Processing: Implications for Morphology Learning. *Lingue e Linguaggio* X(2). 209-226.

Gathercole, S.E., G.J. Hitch, E.S. Service & A.J. Martin (1997). Phonological short term memory and new word learning in children. *Developmental Psychology* 33. 966-979.

Lively, S.E., D.B. Pisoni, S.D. Goldinger & M.A. Gernsbacher (1994). Spoken word recognition: Research and theory. In M.A. Gernsbacher (ed.), *Handbook of psycholinguistics*, 265-301. San Diego: Academic Press.

Luce, P.A. (1986). A computational analysis of uniqueness points in auditory word recognition. *Perception & Psychophysics* 39. 155-158.

Luce, P.A. & D.B. Pisoni (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing* 19. 1-36.

Kullback, S. (1987). Letter to the editor: The Kullback-Leibler distance. *The American Statistician* 41(4). 340-341.

Kuperman, V., R. Bertram & R.H. Baayen (2010). Processing trade-offs in the reading of Dutch derived words. *Journal of Memory and Language* 62(2). 83-97.

Manning, C.D. & H. Schütze (1999). *Foundations of statistical natural language processing*. MIT press.

Marzi, C., Ferro, M. & Pirrelli, V. (2014). Morphological structure through lexical parsability. *Lingue e Linguaggio* XIII(2). 263-290.

Marzi, C., M. Ferro, F.A. Cardillo & V. Pirrelli (2016). Effects of frequency and regularity in an integrative model of word storage and processing. *Italian Journal of Linguistics* 28(1). 79-114.

Marzi, C., M. Ferro, O. Nahli, P. Belik, S. Bompolas & V. Pirrelli (2018). Evaluating Inflectional Complexity Crosslinguistically: a Processing Perspective. In *Proceedings of 11th LREC 2018*, Miyazaki, Japan. Paper 745.

Milin, P., L.B. Feldman, M. Ramscar, P. Hendrix & R.H. Baayen (2017). Discrimination in lexical decision. *PloS one* 12(2). e0171935.

Milin, P., D.F. Đurđević & F. Moscoso del Prado Martín (2009a). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language* 60(1). 50-64.

Milin, P., V. Kuperman, A. Kostić & R.H. Baayen (2009b). Words and paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation. In J.P. Blevins & J. Blevins (eds.) *Analogy in grammar: Form and acquisition*, 214-252. Oxford: Oxford University Press.

Mulder, K., T. Dijkstra, R. Schreuder & R.H. Baayen (2014). Effects of primary and secondary morphological family size in monolingual and bilingual word processing. *Journal of Memory and Language* 72. 59-84.

Orsolini, M. & W. Marslen-Wilson (1997). Universals in morphological representation: Evidence from Italian. *Language and Cognitive Processes* 12. 1-47.

Pirrelli, V., M. Ferro & C. Marzi (2015). Computational complexity of abstractive morphology. In M. Baerman, D. Brown & G. Corbett (eds.), *Understanding and Measuring Morphological Complexity*, 141-166. Oxford: Oxford University Press.

Pitt, M.A. & J.M. McQueen (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language* 39(3). 347-370.

Ramscar, M., M. Dye & J. Klein (2013). Children value informativity over logic in word learning. *Psychological Science* 24(6). 1017-1023.

Ramscar, M. & D. Yarlett (2007). Linguistic Self-Correction in the Absence of Feedback: A New Approach to the Logical Problem of Language Acquisition. *Cognitive Science* 31. 927-960.

Rescorla, R.A. & A.R. Wagner (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory* 2. 64-99.

Vitevitch, M.S. & P.A. Luce (1998). When words compete: Levels of processing in spoken word recognition. *Psychological Science* 9. 325-329.

Vitevitch, M.S., P.A. Luce, J. Charles-Luce & D. Kemmerer (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech* 40. 47-62.

*Marcello Ferro*
Institute for Computational Linguistics (ILC-CNR)
Area della Ricerca, Via G. Moruzzi 1, 56124 Pisa - Italia
email: marcello.ferro@ilc.cnr.it

*Claudia Marzi*
Institute for Computational Linguistics (ILC-CNR)
Area della Ricerca, Via G. Moruzzi 1, 56124 Pisa - Italia
email: claudia.marzi@ilc.cnr.it

*Vito Pirrelli*
Institute for Computational Linguistics (ILC-CNR)
Area della Ricerca, Via G. Moruzzi 1, 56124 Pisa - Italia
email: vito.pirrelli@ilc.cnr.it