SCUOLA
NORMALE
SUPERIORE

laboratorio di linguistica

Basilio Calderone[*], Chiara Celata & Fabio Montermini[*]

# Phonological detail to access morphological structures

*(extended version of the talk given at 7th Décembrettes, Colloque International de Morphologie, Toulouse, December, 2-3, 2010)*

## 1 Introduction

As many other Indo-European languages, particularly of the fusional type, Italian shows bound inflectional morphemes predominantly inserted by suffixation; most grammatical relations and relational categories are overtly expressed by morphological endings more often than by other types of affixes. Consequently, bound function morphemes tend to occupy the right edge of the word. From a quantitative point of view, a contrast between the left and the right edge of a word may trivially be set up by the different statistical properties of morphemes that tend to occur in either position of the word. One and the same phonological sequence will define a set of different quantitative properties (absolute 'token' frequency, frequency of the lexical forms in which it appears, number of neighbors etc.) depending on its position in the word (Table 1). These differences will necessarily impact over lexical processing altogether.

| *ATO #* | | *ATO#* | |
|---|---|---|---|
| nitrato | *NITRATE. m.s. 'nitrate'* | **ato**mico | *(adj.) 'atomic'* |
| mangi**ato** | *EAT.p.part. 'eaten'* | **ato**ssico | *(adj.) 'non-toxic'* |
| am**ato** | *LOVE.p.part. 'loved'* | **ato**llo | *(n.sg.) 'atoll'* |
| pag**ato** | *PAY.p.part. 'paid'* | | |
| bevu**to** | *DRINK.p.part. 'drunk'* | | |
| pos**to** | *PUT.p.part. 'put'* | | |

*Table 1: Example of positional regularities in morphologically complex words: initial vs. final /ato/ in Italian (and grammatically related sequences)*

[*] CLLE-ERSS CNRS & Université de Toulouse Le Mirail.

What does this imply for the theory of lexical access of morphologically complex words?

Within the most accredited models of morpholexical processing, morphemes are not recognized in isolation but rather relationally in the context of other phonologically similar material (Luce et al. 1990, among others). In the non-semantically-driven component of the morpheme recognition process, units in the mind result from contrast, and contrast derives from distributional diversity (e.g., Baayen 2003, Libben & Jarema 2004). Morpholexical processing is in fact affected to a great extent by the statistic properties of the lexicon (or sub-parts of the lexicon), and primarily by quantitative properties of affixes, such as their relative frequency, phonological (and orthographic) neighborhood density, family size, family frequency, and possibly other (e.g., Schreuder & Baayen 1997, Baayen 2003 etc.).

For an inflecting language such as Italian, morpholexical *routines* based on the distributional diversity of morphemes (both across affixes and across roots) have been repeatedly found to be an efficient and frequently activated processing strategy for both word recognition and – more recently – word naming (see Burani & Laudanna 2003 for a recent review).

In previous experiments, we explored the quantitative aspects of morpheme competition in terms of their *positional correlates* in Italian word structure. We asked how positional variables (beside quantitative ones) are processed in blind-to-semantics decomposition of complex words, and wondered whether they represent psychologically and computationally salient pre-conditions for morphological parsing in Italian (Calderone et al. 2008, Celata & Calderone 2011). The results showed that the salience of the right edge of morphological complex pseudo-words (i.e., the portion usually occupied by function elements) emerged as a by-product of micro-phonotactic preferences and macro-phonotactic positional information. By micro-phonotactics we meant sequential information among segments (e.g., the fact that, in the specific language, a phonological sequence such /ato/differs from similar sequences such as /tao/, /rto/, /atu/ etc.). By macro-phonotactics, on the other hand, we referred to positional information within the word, i.e., sub-lexical (or chunk) frequency effects (e.g., the fact that word-initial /#ato/ is different from word-medial /-ato-/ and from word-final /ato#/). This hypothesis was tested on a behavioral and a computational ground, within an experimental protocol aimed at measuring the correlation degree between the speakers' responses and the computational output obtained over one and the same linguistic data set. In particular, morphologically complex pseudo-words were used to elicit ortho-phonological similarity values from both native Italian subjects and an activation-based computational model trained with a phonologically encoded corpus of spoken Italian.

The present experiment aims at replicating the previous findings under the following conditions: by testing the role of lexical stress on morphological decomposition (which implies, among other things, a focus on whole word processing); by evaluating the emergent nature of morphological entities on the basis of a *continuum* of distributional properties.

## 2 Behavioral experiment

### 2.1 Materials

#### 2.1.1 Pseudo-affixes and the distributional *continuum*

As a first step in the realization of the experimental corpus, the quantitative properties of 35 Italian prefixes and suffixes were analyzed according to the following procedures. Basing on a large list of derivational affixes taken from Gradit (De Mauro et al. 1999-2007), 21 prefixes (of which 16 disyllabic and 6 monosyllabic) and 14 suffixes (of which 12 disyllabic and only one monosyllabic) were selected. For each of them, the following data were drawn from CoLFIS (Bertinetto et al. 2005), a large and representative lexical data-base for written Italian (3.798.275 lexical entries) available online: (i) number of word forms (i.e., lexical contexts) in which the phonological sequence corresponding to the selected affix was present in initial position; (ii) number of word forms (i.e., lexical contexts) in which the phonological sequence corresponding to the selected affix was present in final position; (iii) number of token forms (i.e., lexical contexts considered with their actual frequency of occurrence) in which the phonological sequence corresponding to the selected affix was present in initial position; (iv) number of token forms (i.e., lexical contexts considered with their actual frequency of occurrence) in which the phonological sequence corresponding to the selected affix was present in final position. It is worth noticing that these quantitative data were calculated with respect to the phonological sequence corresponding to the real affix, and not to the affix itself; so for example, when the prefix *ambi-* was selected, CoLFIS was asked to provide the number of word forms/tokens beginning by *ambi-* in cases (i) and (ii) (e.g., *ambisillabico* 'ambisyllabic' but also *ambizione* 'ambition'), and the number of word forms/tokens ending by -*ambi* in cases (iii) and (iv) (e.g., *giambi* 'iambs'). The rationale behind the choice of including non-affixed forms such as *ambizione* (as well as of calculating the frequency of a sequence phonologically coincident with a prefix even when occurring in final position, such as *ambi* in *giambi*, and vice-versa) will emerge straightforwardly in the following sections. For the moment, it is important to point out that, in the continuation of the paper, we will be referring to this empirically defined notion of sequence phonologically coincident with true Italian affixes by means of the term 'pseudo-affix' (in particular, 'pseudo-prefix' and 'pseudo-suffix').

This list of 35 pseudo-affixes, outfitted with distributional information on word-initial and word-final occurrences, constituted therefore the initial database for subsequent pseudo-affix selection and non-word realization.

The relation between the distributional behavior of pseudo-affixes and the expected behavior of the phonologically coincident real Italian affixes could vary across pseudo-affixes. In the majority of the cases, the frequency of occurrence showed by pseudo-affixes in word-initial vs. word-final position was totally consistent with the expected behavior of the corresponding Italian affix. This kind of relation is exemplified, in Table 2, by the sequence *pre*, which coincides with a real prefix in Italian, and whose frequency of occurrence in word-initial position is consistently higher than in word-final position (for both word forms and token forms). This condition was considered to be distributionally unambiguous and there was a consistent relation between the grammatical status of the affix and the rough distributional information associated to the phonological sequence. In a second class of pseudo-affixes, the frequency of occurrence in word-initial vs. word-final position was not consistent with the expected behavior of the corresponding Italian affix; a clear example for this pattern can be seen in the sequence *rico*, which coincides with an Italian suffix but, if one looks at the CoLFIS data, turns out to be much more represented as a word-initial than word-final sequence, especially when token forms are considered. This second condition was considered to be distributionally unambiguous (no mismatch between word forms and token forms data) but there was no consistency between the grammatical status of the affix and the rough distributional information associated to the phonological sequence. Finally, the corpus also included pseudo-affixes unevenly attested in terms of word forms vs. token forms count: for example, the sequence *iso* (which coincides with an Italian prefix) resulted to be more represented in word initial than word final position when word forms were considered, but the reverse pattern was found when word tokens were looked at. We interpreted this latter condition as one of distributional ambiguity (because of a mismatch between word forms and token forms data) coupled with an inconsistency in the relation between the grammatical status of the affix and the rough distributional information associated to the phonological sequence (because of contrastive evidence coming from the two sources of data).

In sum, by considering the quantitative properties of pseudo-affixes in terms of both word forms and token forms and by crossing the two parameters of distributional ambiguity and consistency between distributional data of pseudo-affixes and grammatical status of homophonous affixes we traced the distributional continuum back to a grid of interpretable conditions which were used empirically in the experimental analysis to be presented.

| Phonemic sequence | Word forms Initial | Word forms Final | Token forms Initial | Token forms Final | Grammatical status of the affix | Distributional ambiguity | Consistency distribution / grammatical status |
|---|---|---|---|---|---|---|---|
| pre | 1214 | 12 | 20.638 | 4.752 | prefix | unambiguous | consistent |
| rico | 358 | 102 | 3.938 | 1.428 | suffix | unambiguous | inconsistent |
| iso | 57 | 40 | 684 | 2.360 | prefix | ambiguous | inconsistent |

*Table 2: Example of pseudo-affixes and their ditributional properties (source: CoLFIS, Bertinetto et al. 2005)*

### 2.1.2 Pseudo-words and association conditions among them

A set of pseudo-words was realized by combining each pseudo-affix with phonotactically legal non-roots (see Table 3). Each pseudo-affix was included in two pivot items, one in which the sequence was placed in initial position (e.g., ***pre**luma*), and the other in which the sequence was placed in final position (e.g., *mula**pre***). From a segmental point of view, the two pivot items were exactly the same; only the relative position of segments was different. Each pivot item contrasted with two associated items, where the same pseudo-affix was combined with a different non-root. In the first associate, the pseudo-affix shared the same position than the pivot (e.g., for a pivot ***pre**luma*, an associated item ***pre**niso* was created, and for a pivot *mula**pre***, an associated item *soni**pre*** was created). In the second associate, the pseudo-affix was split up within the word (e.g., *pornesi* was associated to *mulapre*, and *sornepi* to *preluma*). It has to be noted that the two associates of each set were exactly equivalent to each other with respect to the segmental composition, but different to the extent that the pseudo-affix could be placed in either the same, or a different position with respect to the pseudo-affix contained in the pivot. The initial and final segments of both associates in the non-identity condition (e.g., *sornepi* and *pornesi*) were different from the initial and final segments of the corresponding pivot items. It is also worth noticing that, since each pseudo-affix was placed in both initial and final position of the morphologically complex pseudo-word, it turned out to appear in its 'licit' (e.g., *preluma*, *preniso*) as well as 'illicit' (*mulapre*, *sonipre*) position. Therefore, pivots and type 1 associates made up of pseudo-prefixes in the initial positional option can be considered morphological pseudo-words in the typical sense (e.g., Burani et al. 1995; Laudanna, Cermele & Caramazza 1997), while when they occur in the final positional option (e.g., *mulapre*, *sonipre*) the level of decomposability is undoubtedly inferior (whether it exists at all is exactly what is at issue here) and they can provisionally be considered 'pseudo-morphological pseudo-words'. The reverse will be true for pivots and type 1 associates made up of pseudo-suffixes: they can be considered morphological pseudo-words when they occur in the final positional option, pseudo-morphological pseudo-words  in the initial positional option. Each pseudo-affix was therefore associated to morphological and pseudo-morphological pseudo-words in the same proportion.

| | | | Positional options | |
|---|---|---|---|---|
| | | | Initial position | Final position |
| | | Pivot | **pre**luma | mula**pre** |
| Identity conditions | Identity | Associate - type 1 | **pre**niso | soni**pre** |
| | Non-Identity | Associate - type 2 | so**r**n**ep**i | **p**o**r**n**e**si |

*Table 3: Example of pseudo-words*

Stress position was carefully controlled for. Stress was always on the penultimate syllable of the pseudo-words. Stress position was balanced across the identity/non-identity conditions, with respect to the quality of the stressed vowel; for example, if the stressed vowel was different in the first associate with respect to the pivot (such as in the *preluma* – *preniso* example above), it was also different in the second associate with respect to the pivot (e.g., *preluma* – *sornepi*). In addition, disyllabic pseudo-words were consistently stressed or unstressed across each members of a pair, in such a way that the two associate conditions and the two positional options did not differ with respect to this parameter. For example, in the following series: (i) *arolanti-evipanti*, (ii) *arolanti-ivintape*, (iii) *antiralo-antivepi*, (iv) *antiralo-ivintepa*, the stressed vowel was /a/ for both members of the pairs (i) and (ii), while there was an alternation between /a/ and /e/ across the members of the pairs (iii) and (iv).

By controlling for such segmental and supra-segmental regularities across the relevant pairs of pseudo-words, we wanted to warrant perfectly balanced conditions of phonological variation across identity conditions and positional options.

A list of 140 (35 pseudo-affixes × 2 positional options × 2 identity conditions) experimental pairs was finally created.

### 2.2 Procedure and analysis

A pool of 22 native Italian subjects (aged 25-35) was asked to judge the similarity of each pivot item with respect to either one of the two associated items. 16 pairs (some of them reproducing the pattern of alternation described above for the experimental pairs, and some others made of two identical non-words, such as *muserota-muserota*) were created and used as fillers. The resulting 164 pairs were randomly disposed in four blocks of 41 pairs such that each block contained one-fourth of the pairs in each category and participants saw all four conditions for each pseudo-affix across blocks.

A beep sound signaled the beginning of the trial. The stimuli were visually and aurally presented. The subjects were sitting in front of a computer screen, where one list was presented at a time, and they were wearing headphones, through which they could hear a female native Italian speaker pronouncing each pseudo-word pairs with an inter-stimulus interval of 100 msec and an inter-trial interval of 3500 msec. The beginning of each trial was signaled by a beep sound notifying the appearance of a new trial. The subjects were asked to judge the degree of formal similarity between the members of each pair with respect to a 9-point scale (1 = minimum similarity, 9 = maximum similarity or identity) and write the corresponding value in a cell on the right of the written stimulus. The subjects were previously instructed about the range of phonological variation they will have been coming across in the experiment, and in particular, they were told that the members of each pair did not differ for length and stress position, and that some pairs were constituted of identical stimuli (whose expected similarity judgment was 9). They were then asked to exploit the whole range of the scale in giving their judgments and modulate the use of the numerical values according to the global range of phonological variation attested in the experiment. A familiarization phase preliminary to testing was added with the specific purpose of allowing individual judgments calibration. The subjects were also informed about the totally offline nature of the experiment, which would have allowed them to make cross-trial comparisons and even auto-corrections all along the experimental session. Finally, they were reassured on the absolute subjectivity of the judgments and encouraged to rely on their own individual strategy for task completion. Each subject performed the test individually and the experimental session lasted 20 minutes overall.

The similarity ratings given by the subjects were treated as dependent variables and evaluated through an analysis of variance. In our hypothesis, the identity condition should elicit higher similarity values than the non-identity condition in the final positional option more than in the initial one. For example, we expect *preluma* to be judged much more similar to *preniso* than to *sornepi*, yet the pair *mulapre* - *sonipre* to be judged more similar than the pair *mulapre* - *pornesi* only to a lesser extent. The perceived difference between the identity and the non-identity conditions is therefore expected to be greater for stimuli containing pseudo-affixes in final position than for those containing pseudo-affixes in initial position (irrespective of the grammatical nature of the affix). In statistical terms, we expect a significant interaction between the two independent variables of identity condition and positional options, uniformly for pseudo-words made of non-root + pseudo-prefixes and of non-root + pseudo-suffixes.

### 2.3 Results

Univariate ANOVAs were run with identity condition (identity vs. non-identity) and positional option (initial vs. final) as between-subject factors. Overall, there were significant main effects of identity condition (with higher values for identity than for non-identity; F (1, 3079) = 3677,351, p < .001) and positional option (with higher values for final position than for initial position; F (1, 3079) = 26,543, p < .001); the interaction of the two was found to be statistically significant (F (3,3079) = 32,744, p < .001), thus indicating that the difference between 'identical' and 'non identical' pairs of stimuli was differently perceived by the subjects, depending on the relative position of pseudo-affixes in the pivots. In particular, as the average values clearly showed, there was a stronger effect of identity for the subset of stimuli in the final positional option (identity: 5,88; non-identity: 2,18) than for those in the initial positional option (identity: 5,14; non-identity: 2,24).[1]

Recall that the experimental pairs were perfectly balanced with respect to phonological variables, including stress position (§2.1). The same pseudo-affixes were used to create pseudo-words for both the final and the initial positional options, and the rules of anagrammatizing were kept constant across each stimulus pair. In principle, if the subjects had been producing similarity judgments by simply calculating the number of shared and non-shared segments between the two members of each pair, their responses would not have differed for any of the experimental subsets. It must then be concluded that any variation in the perceived similarity of different subsets of stimuli has to be connected to factors unrelated to segmental constituency of the stimuli and phonology in general. For us, the present results clearly indicate that, all other things being equal, the presence of phonological regularities in word final position is more salient than the presence of the same regularities in word initial position. Given that not only pseudo-suffixes, but also pseudo-prefixes could occur in final position, it has to be pointed out that such alleged phonological regularities are of a very specific nature: they are chunks of segments defined in terms of bare co-occurrence in the corpus, with respect to specific (either word final or word initial) positions, and independent of the functional or grammatical status of phonologically similar strings. As we have already illustrated, our pseudo-affixes stand out in purely distributional terms as the by-product of their relative frequency of occurrence in different positions within the word; their salience in the similarity task uniquely derives from these distributional characteristics.

Therefore, the present results support the hypothesis of a differential salience of the right edge of the word in similarity ratings produced by native speakers of Italian; the final portion of the word thus emerges as a position with a great potential for lexical processing altogether. If this hypothesis is correct, the distributional properties of phonological strings (also referred to above as the cumulative effect of micro- and macro-phonotactic regularities) should be considered as outstanding preconditions of morphological parsing.

[1] The same effect held strong independently of the grammatical status of pseudo-affixes (as pseudo-suffixes or pseudo-prefixes), as shown by the non-significant interaction position*identity*affix (F (7, 3079) = 0,110, p > .05).

## 3 Computational simulation

The computational simulation was grounded in PHACTS (for *PHonotactic ACTivation System*), a computational model of how ortho-phonological words are learned, processed and retrieved in the mental lexicon of the speakers (Calderone et al. 2008, Celata & Calderone 2011). The same set of morphologically complex pseudo-words described in §2 was used to elicit similarity values from a topological neuronal receptive map trained with a phonologically encoded corpus of written Italian (Quasthoff et al. 2006, Italian section).

PHACTS is based on the principles of a Self-Organizing Map (SOM), which is a neurocomputing algorithm for multivariate analysis (Kohonen 2001). PHACTS simulates the formation of phonotactic knowledge in the mind of a speaker, who is exposed to a stream of phonological words and gradually reaches a knowledge representation of the statistical regularities shaping the phonotactics of a given language. Starting from this kind of statistical knowledge, the model is able to generalize the phonotactic information to novel stimuli deriving activation-based representations of full lexical forms, judging the well-formedness of unseen stimuli or even attributing salience to a specific word subpart disregarding the others.

As a consequence, PHACTS outputs multi-dimensional word representations from phonotactic knowledge appealing solely to local conditions on language-specific phonotactic constraints.

### 3.1 PHACTS: the algorithm

The physical structure of PHACTS is defined by a set $S$ (with cardinality of finite sets) of neurons $n_{jk}$ with $1 \leq j \leq J$ and $1 \leq k \leq K$ arranged in a bi-dimensional grid of $S = \{n_{11}, n_{12}, \mathrm{K}\ n_{JK}\}$, |S| = JxK. Each neuron $n_{jk}$ in the grid corresponds to a vector $u_{jk}$ (the so-called prototype vector) whose dimension is equal to the dimension of the data vector $i$ that will be the input to the system. Before the learning phase, the prototype vectors $n_{jk}$ assume random values and during the learning phase they change these values in order to adapt themselves to the input data $i$.

PHACTS operates following two successive phases: a learning phase, and a transfer function phase.

a) The learning phase: the search for the BMU and the topological adaptation

Before the learning phase each input is presented iteratively to the model. At each iteration the algorithm searches for the *best matching unit* (BMU), that is the neuron topologically closer to the input vector $i$ and which is candidate to represent the input data through the prototype vector. The search for the BMU is given by maximizing the dot product of $i$ and $u_{jk}$ in the *t*-th in the step of the iteration:

$$BMU((i)t) = \arg\max_{jk}(i(t) \cdot u_{jk}) \tag{1}$$

In other terms, the $BMU((i)t)$ is the prototype vector better aligned with the input $i$.

After the $BMU$ is selected for each $i$ at time $t$, PHACTS adapts the prototype vector $u_{jk}$ to the current input according to a topological adaptation equation given in (2):

$$\Delta u_{jk}(t) = \alpha(t)\delta(t)[i(t) - u_{jk}(t-1)] \tag{2}$$

where $\alpha(t)$ is a *learning rate* and $\delta(t)$ is the so-called *neighborhood function*. The *neighborhood function* is a function of time and distance between the $BMU$ and each of its neighbors on the bi-dimensional grid. In other terms it defines a set of nodes around the $BMU$ that would receive training, while nodes outside this set would not be changed. In our model the *neighborhood function* is defined as a Gaussian, where $2\sigma^2$ is a value of distance between the $BMU$ and its neighboring neurons:

$$\delta(t) = \exp\left(-\frac{\left\|u_{jk}(t-1) - BMU(i(t))\right\|^2}{2\sigma^2}\right) \tag{3}$$

The learning rate parameter controls for the elasticity of the network, and the neighborhood function roughly controls for the area around each best matching where the neurons are modified. The initial value of both parameters is set heuristically and in general decreases as long as the learning progresses. In order to facilitate a training convergence, we set $\alpha \to 0$ and $\delta \to 0$ as $t \to \infty$.

PHACTS performs a vector mapping of the data space in input to the output space defined by the prototype vectors $u_{jk}$ and the bi-dimensional position onto the grid of $S$ neurons.

After the learning, each input $i$ occupies two positions: one in the output space (through the prototype vectors) and one in the bi-dimensional space (defined by specific coordinates of the grid).

Due to $BMU$ and to the topological adaptation equation in (2), the more frequent the input stimuli of the training corpus are (in terms of token frequency), the greater the activation of the corresponding neurons. This frequency effect is reflected also in the transfer function phase, as explained below.

b) Transfer function

Once PHACTS has been trained by exposition to input vectors (we will specify the nature of the input and its linguistic motivation in the next section), an activation-based representation of unseen stimuli can be derived with respect to the output space previously obtained. This phase implements a linear thresholded function in which each neurons ‘fires’ as a function of its activation with respect to the unseen input. In this sense each neuron acts as a ‘transfer function’ of an activation weight depending on the alignment between the unseen input vector and the prototype vector.

Let $x$ be an input vector (not present in the training set) and $u_{jk}$ the prototype vector of the neuron $n_{jk}$ and let $d$ a threshold value (heuristically set). Let $\Phi_{jk}$ be the transfer function of the neuron $n_{jk}$ defined as in (4):

$$\Phi_{jk}(x)=\max\left(\left(0, x\cdot u_{jk}\right)-d\right) \qquad (4)$$

where $\Phi_{jk}(x)\geq 0$.

As a consequence, the activation-based representation of input $x$ defines a sort of distributed representation of the input $x$ which reflects both the position in the grid’s bi-dimensional topological organization and the activation of the neuronal output space previously trained.

## 3.2 Present simulation

In this experiment, PHACTS has been implemented to derive activation-based representations of phonological words from an output space trained with a phonologically transcribed corpus of written Italian. In the following sections the details of the experiment are provided.

a) Learning phase: the creation of a phonotactic knowledge for Italian

A corpus of written Italian from the *Leipzig Corpora Collection* (Quasthoff et al. 2006) was used for the training phase. The corpus is phonologically transcribed and contains nearly 80,000 word types and 5 millions word tokens. The size of the bi-dimensional grid was 25 X 35 neurons so that S = 875.

Differently from previous experiments (Calderone et al. 2008, Celata & Calderone 2011), where consonants were specified for place, manner of articulation and voicing and vowels were specified for roundedness, height and anteriority, we wanted here to include specific coding for lexical stress. Given the distinctive value of lexical stress in Italian, information about the position of lexical stress would provide the system with a more detailed representation of word-sized input, thus mirroring the native speakers' representation to a larger extent. In the present experiment, therefore, vowels were additionally specified for a feature referring to their prominence in the word ([ ± stress]).

It should be noted that stress notoriously is a property of (phonological) words, not just of syllables or, even less so, individual vowels. Adding a binary feature to encode the prosodic prominence of a syllable within the word is but a rough approximation of the supra-segmental dimension of lexical stress. In terms of coding, the question is open to future research and the present experiment should therefore be considered as explorative with respect to the possibility of encoding suprasegmentals in the phonological equipment of the system.

In this experiment, string sampling was fixed at n = 6 (hexagrams). Setting a sampling window is actually a matter of empirical preferences and the choice is justified because of the need of finding a heuristically defined compromise between the phonotactic (string level) and the lexical (word level) knowledge representation; the optimal value therefore depends on a series of independent conditionings such as the type and length of the stimuli, the phonotactics of the language under investigation, as well as the dimension of the map and the number of involved neurons, and possibly many others. Common practice is to use bi- or tri-gram sampling windows to recover phonotactic local constraints in the input data, and we also did use tri-gram sampling in previous experimentation with PHACTS (Calderone et al. 2008, Celata & Calderone 2011). When a larger sampling window is used (and independently from other relevant factors such as the dimension of the map and the quantity of the input), the neurons specialize on larger units and therefore regularities involving a greater number of phonemes will be recovered as more salient than when bi- or tri-gram windows are used (while regularities involving few segments will produce smaller effects since they are diluted on larger representation windows). Given some general, large-scale characteristics of the Italian lexicon ('average' word length, ratio of unstressed to stressed syllables per word etc.), we decided to use a 6-gram sampling window in order to derive a sufficiently sharp representation of regularities involving relatively large chunks of segments, thus simulating a full-word memorization process where both segmental and supra-segmental variables are taken into in account.

Each word of the corpus therefore passed a 6-grams sampling and in this form was given in input to the system. As explained, as a consequence of the 6-grams sampling the model develops a topological organization of the stimuli that reflects a sort of phonotactic knowledge enriched with quantitative information on token frequencies: similar *n*-gram tokens are mapped adjacent one to another onto the bi-dimensional grid and high-frequency *n*-grams exhibit a stronger activation degree at the level of the transfer function, compared to low-frequency *n*-grams.

b) Transfer function: activation-based representations of pseudo-words

On the basis of the topological organization obtained in the learning phase, each neuron of the bi-dimensional grid acts as a 'transfer function' for unseen input items. Specifically, the pseudo-words used in the behavioral experiment (§2.1) were given in input to the algorithm after the training was completed. To obtain a final vector representation of the word, the system performed a transfer process by summing the activation values of each 6-gram $x$ as stated in the equation in (4) above; see (5):

$$F_{PHACTS}(x)=\sum_{jk}\Phi(x) \qquad (5)$$

The cumulative action of n-gram activations gave a graded and distributed representation of the word in which both phonological similarity (at the string level) and token frequency effects (at the level of words and sub-lexical structures) were taken into account.

The final activation-based representation of a pseudo-word was hence a vector representation of 875 dimensions (recall the number of neurons in the map was $S = 875$), as represented in Fig. 1.

Such granular and enriched representation allows us straightforwardly to compare pseudo-words each other and derive a measure of their formal similarity. This was achieved by adopting the cosine values of the angle between two vector representations. Consequently, the similarity scale ranged from 1 (total identity) to 0 (total dissimilarity).

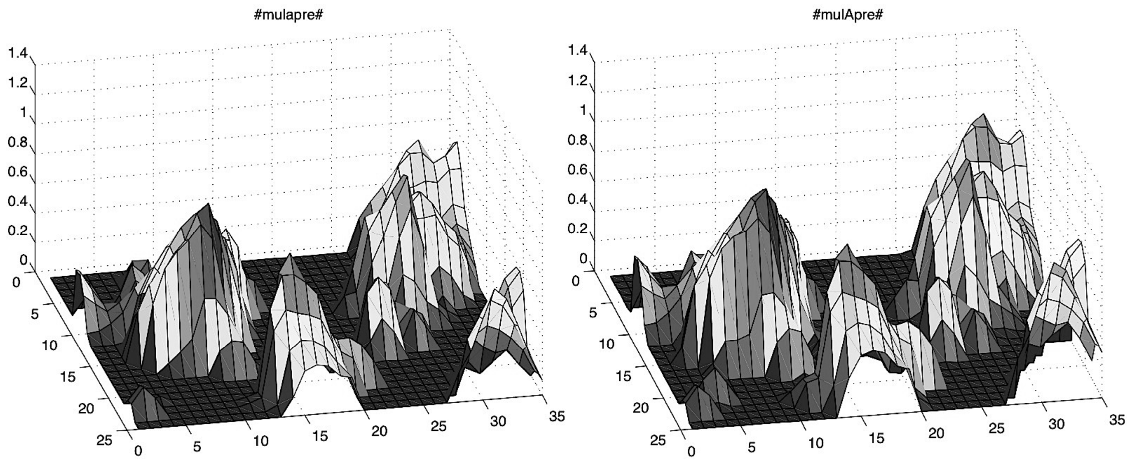

*Fig. 1. Comparison between the activation-based representations of the pseudo-word #mulapre#, with no stress specification (on the left) and #mulApre#, with the vowel of the second syllable specified as [+ stress] (on the right)*

### 3.3 Results

The cosine values given in output by the system were taken as the independent variable and the effects of the factors of identity condition (identity vs. non-identity) and positional option (initial vs. final) were analyzed as between-subject factors.

As expected, there was a significant main effects of identity condition (with higher values for identity than non-identity; F (1,139) = 56,122, p < .001) while the effect of positional option was non-significant overall (F(1, 139) = 0,930, p > .50); the interaction of the two was found to be statistically non-significant (F(3, 139) = 1,110, p > .50). The average values actually showed that there was a stronger effect of identity for the subset of stimuli in the final positional option (identity: 0,92 *vs* non-identity: 0,61) than for those in the initial positional option (identity: 0,84 *vs* non-identity: 0,62), but the difference did not reach the level of statistical significance, possibly because of relatively high standard deviation values (0,057 *vs* 0,304 in final position; 0,149 *vs* 0,253 in initial position).

Given the numerical dispersion associated to variability of the system's output, we wanted to explore the possibility that the expected interaction, indicating - all other things being equal - a different salience for word-final similarities with respect to non-word-final similarities, did hold true for some empirically defined subset of the data. Our first hypothesis was that the system was likely to be sensitive to the distributional information of the input in a more straightforward way compared to native speakers, possibly suffering the consequences of statistical drifts inherent to the data. We therefore wanted to check whether different types of affixes elicited different responses from the system. The parameters of distributional ambiguity and of consistency between distributional data and grammatical status (see above, §2.1) were thus used as possible predictors of the system's reactions. The latter parameter was not found to influence the response pattern, since both consistent and inconsistent pseudo-affixes elicited a non-significant position*identity interaction. On the contrary, distributional ambiguity did prove to be the relevant parameter, since only pseudo-words made of distributionally unambiguous pseudo-affixes exhibited a significant position*identity interaction (F (3,123) = 12,110, p < .05), with a stronger effect of identity for items in the final positional option (identity: 0,92 *vs* non-identity: 0,59) than for items in the initial positional option (identity: 0,83 *vs* non-identity: 0,63). Recall that by distributional ambiguity we indicated the condition of those pseudo-affixes unevenly attested in terms of word forms *vs* token forms count (§2.1); given this definition, distributional ambiguity also implies inconsistency between distributional data and grammatical status of the corresponding affix (all ambiguous pseudo-affixes also showed inconsistency between statistical distribution and the grammatical status of the homophonous affix, but the reverse was not true; see above, §2.1). We can consequently conclude that, when there is no mismatch between different sources of quantitative information relative to the training data, PHACTS is able to let the salience of the right edge of the word emerge, in much the same way than the native speakers of Italian unconditionally do.

A second non-exclusive hypothesis was put forward, referring to a possible role of phonotactic constituency for pseudo-affixes. Specifically, it was hypothesized that pseudo-affixes composed of more segmental units were more likely to be autonomously parsed by the system with respect to short pseudo-affixes; consequently, they were supposed to facilitate the system in generalizing the information about the salience of the final portion of the word. Phonological length, calculated in terms of number of syllables composing the pseudo-affix included in the experimental items, was therefore considered as a further predictor of the system's reactions to the similarity evaluation task. The length factor thus entered the analysis to get a picture of whether the items composed of monosyllabic pseudo-affixes and the items composed of disyllabic pseudo-affixes performed differently with respect to the expected positional effect. The hypothesis however turned out to be disconfirmed by the data: identity*position*length was non-significant (F (7, 139) = 0,472, p > .05) and the two subsets of experimental items elicited a non-significant identity*position interaction, thus indicating that the length of the pseudo-affix is irrelevant for the transfer function operated by PHACTS on novel stimuli, at least as far as the final position salience effect is concerned.

## 4 Human/machine correlation

The Pearson's correlation coefficient between the observed and the simulated behavior (r = 0.604) described a statistically significant correlation (p < .001), thus confirming the psychological plausibility of the SOM-based simulation (Fig. 2).

It should also be noted that higher levels of consistency between the system's behavior (transformed through exponential function) and the speakers' behavior (median of the list of judgments) were obtained for the identity condition (both final and initial position) with respect to the non-identity condition, where much more fluctuations were observed in the system's output and the cosine values covered an area of relatively greater dispersion. The speakers overall appeared to be much more consistent in assigning similarity values to pairs of non-words sharing the same phonemic sequence in different positions, while greater oscillations were produced by the system with respect to this subset of data.

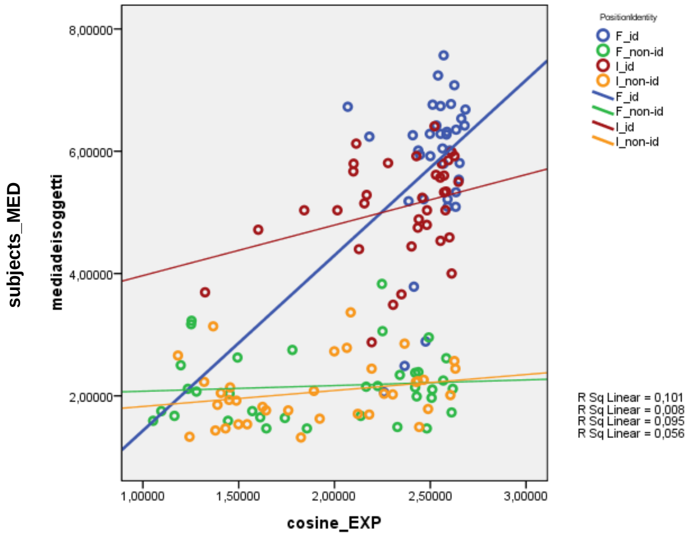

*Fig. 2. Human/machine correlation, split by identity condition and positional option*

## 5 Conclusions

In this study, we replicated the experimental design of previous investigations (Calderone et al. 2008, Celata & Calderone 2011) to the extent that human and artificial data were elicited and compared in response to the same set of Italian pseudo-words, yet some relevant changes were introduced in the domain of the nature of the phonological information coded in the corpus used for PHACTS's training, and the generalization process used to derive the system's word-level representation. The two domains actually were strictly interconnected, inasmuch as a modification in the phonological representation of data required specific amendment in the patterned sampling operated by the algorithm.

The results supported the hypothesis that phonological specifications at the supra-segmental level improve the system's performance in recovering the phonological similarity of stimuli as shaped by positional regularities at the word level, thus providing a more accurate simulation of native speakers' performance on the same task. Differently from the previous experiment, where vowels were unspecified for stress, we found here a significant interaction between identity condition and affix position for the subgroup of unambiguous items, thus confirming the system's ability to recover word-level 'paradigmatic' regularities, besides string-level phonotactic information. Moreover, the Pearson's correlation coefficient changed from r = 0.563

of the previous experiments to r = 0.604, thus indicating that supra-segmental information allowed the system to overlap human generalizations to a larger extent.

The present results have implications for both theories of lexical access and distribution-based models of morphology.

With respect to lexical access, this study demonstrates that competing forces operating at the micro- and macro-phonotactic levels may account for processes of morpheme isolation within the word. Phonological detail (such as lexical stress) enriches morphemic representation in both human and artificial responses to complex words. A 'suffixation preference' (e.g., Hupp et al. 2009) is at work in an inflecting language such as Italian, over and above the assumed position-specific representation of affixes (Crepaldi, Rastle & Davis 2010).

In terms of probabilistic and analogy-based models of morphology, this study has shown that positional variables, besides quantitative properties of (sub)lexical forms, are computationally salient prerequisites for whole-word decomposition. Distributional effects, particularly those related to positional variations at the sub-lexical level, should be kept into account in (interactive, hybrid) models of morpholexical processing. Interactions between 'routes' or 'levels' are organized with respect to morpheme and word specific characteristics which apparently go beyond the effects of well-known quantitative properties such as morphemes' relative frequency, neighborhood density, family size etc. and involving the micro- and macro-phonotactics of the language as an additional (even confounding) source of processing information. Such graded notion of morphological network (which partly reminds of Libben 2003 and similar proposals) subsumes a new concept of formal similarity holding among words and sub-lexical structures and derived from a non-linear interaction of multiple factors affecting the properties of the phonological string (phonemic content, phonotactic regularities, relative frequency of phonemes and of chunks of phonemes) as well as the shape of word-sized units and their mutual relationships in the mental lexicon (positional constraints on the phonemic constituency of words, 'paradigmatic' relations among forms, token frequency distributions).

## Bibliographical references

Baayen H. (2003) Probabilistic approaches to morphology. In R. Bod, J. Hay & S. Jannedy (eds.) *Probabilistic linguistics*. MIT Press, Cambridge MA, 229-287.

Bertinetto P.M., Burani C., Laudanna A., Marconi L., Ratti D., Rolando C., Thornton A.M. (2005) *Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS).* http://linguistica.sns.it/CoLFIS/CoLFIS_home.htm

Burani C. & Laudanna A. (2003). Morpheme-based lexical reading: Evidence from pseudo-word naming. In E. Assink & D. Sandra (Eds.), *Reading complex words: Cross-language studies* (pp. 241-264). Dordrecht: Kluwer.

Burani C., A.M. Thornton, C. Iacobini & A. Laudanna (1995) *Investigating morphological non-words*. In W.U. Dressler & C. Burani (eds.), *Cross-disciplinary approaches to morphology*, Wien, Osterreichischen Akademie der Wissenschaften: 37-53.

Calderone B., C. Celata & I. Herreros (2008) Recovering morphology from local phonotactic constraints. In P. Warren (ed.), *Abstracts of the Laboratory Phonology 11*, Victoria University of Wellington.

Celata C. & B. Calderone (2011) Restrizioni fonotattiche, pattern lessicali e recupero delle regolarità morfologiche. Evidenze computazionali e comportamentali. In *Linguaggio e cervello / Semantica,* Atti del XLII Convegno della Società di Linguistica Italiana (Pisa, Scuola Normale Superiore, 25-25 settembre 2008). Roma: Bulzoni. 47-72. Volume 2 (CD ROM).

Crepaldi D., Rastle D. & C.J. Davis (2010) Morphemes in their place: Evidence for position specific identification of suffixes. *Memory & Cognition* 38: 312-321.

*Gradit- Grande dizionario italiano dell'uso*, ideato e diretto da T. De Mauro, 6 voll.+CD-rom, UTET, Torino 1997-2007.

Hupp J. M., Sloutsky V. M. & Culicover P.W. (2009) Evidence for a domain general mechanism underlying the suffixation preference in language. *Language and Cognitive Processes*, 24, 876-909.

Kohonen T. (2001) *Self-Organizing Maps*. Heidelberg: Springer-Verlag.

Laudanna A., A. Cermele & A. Caramazza (1997) Morpholexical representations in naming, *Language and Cognitive Processes* 12, 49-66.

Libben, G. (2003) Morphological Parsing and Morphological Structure. In A. Egbert & D. Sandra (Eds.) *Reading complex words* (pp. 221-239). Amsterdam: Kluwer.

Libben G. & G. Jarema (2004) Conceptions and questions concerning morphological processing. *Brain & Language* 90: 2-8.

Luce P.A., D. B. Pisoni & S.D. Goldinger (1990) Similarity neighborhoods of spoken words. In G. Altmann (ed.), *Cognitive Models of Speech Processing*, Cambridge, MIT Press: 122-147.

Quasthoff U., Richter M. & Biemann C. (2006) Corpus portal for search in monolingual corpora. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, 1799–1802.

Schreuder R. & R.H.Baayen (1997) How complex simplex words can be. *Journal of Memory and Language* 37: 118-139.