Enriched sublexical representations to access morphological structures. A psycho-computational account

**Chiara Celata*[*] — Basilio Calderone*[**] — Fabio Montermini*[**]

[*] *Laboratorio di Linguistica; Scuola Normale Superiore*
Piazza dei Cavalieri, 7 I-56126 Pisa
c.celata@sns.it

[**] *CLLE-ERSS (UMR 5263) CNRS & Université de Toulouse-Le Mirail*
5, allées Antonio-Machado F-31058 Toulouse Cedex 9
basilio.calderone@univ-tlse2.fr
fabio.montermini@univ-tlse2.fr

ABSTRACT. *The paper investigates the morphological impact of quantitative properties of lexical and sublexical structures in the decomposition of morphologically complex words by means of an activation-based simulation. A complex nucleus of blind-to-semantics relationships turns out to allow the emergence of proto-morphological representations and to provide a cognitively efficient route to complex word decomposition. A computational account of how this bundle of information is handled with in the lexical processing of complex pseudo-words is provided. The model is tested against an edit-distance algorithm and a non-word similarity rating task performed by a group of native speakers.*

RÉSUMÉ. *L'article analyse l'impact morphologique des propriétés quantitatives des structures lexicales et sous-lexicales dans la décomposition de mots morphologiquement complexes en faisant appel à une simulation basée sur des états d'activation. Nous observons qu'il existe un noyau complexe de relations extra-sémantiques capable d'activer des représentations proto-morphologiques et de canaliser efficacement la décomposition des mot complexes. Nous proposons un modèle de manipulation de ce noyau d'information dans l'élaboration lexicale de pseudo-mots complexes. Le modèle est testé par rapport à un algorithme de distance d'édition et à des jugements sur la similarité de mots fournis par des locuteurs natifs.*

KEYWORDS: *Morphological processing, phonotactics, psycho-computational simulations.*

MOTS-CLÉS : *Élaboration morphologique, phonotaxe, simulation psycho-computationnelle.*

1. Introduction

In neurally inspired models of morphological effects in word recognition and processing, the units of morphological decomposition are not primitive objects that explicitly constitute a morphological level of representation, but they are derived in the course of learning from statistical regularities that converge to identify a set of processing units in the mapping between form and meaning. Since Rumelhart & McClelland's model of English past tense production (Rumelhart and McClelland, 1986), morphological "rules" (together with their "exceptions") have been replaced by the dynamics of a computational mechanism basically grounded in the quantitative description of the language. As a consequence, morphological structure has become the preferential window on the cognitive representation of ortho-phonological words and has motivated a number of different experimental investigations into the psycho-computational domain.

For example, extensive evidence has been produced suggesting that, in word reading, the representations of constituent elements serve as access units to the mental lexicon via a pre-lexical semantically blind process that extracts them from the visual input whenever certain conditions are met (Taft, 1994; Taft, 1979; Caramazza *et al.*, 1988; Schreuder and Baayen, 1995; Meunier and Segui, 2002; Rastle and Davis, 2004); morphological decomposition takes place at a supra-lexical level of analysis following slightly different hypotheses (Giraudo and Grainger, 2003). Moreover, computational modeling studies have addressed the phenomena related to morphological effects in word recognition by simulating the network's sensitivity to morphological regularities in associating specific phonological features (such as "final -s" or "final -ed" in English) with appropriate semantic features (such as "plural"/"third person singular" and "past tense") (Harm and Seidenberg, 2004; Rueckl and Raveh, 1999; Plaut and Gonnerman, 2000).

Taken together, these psycho-computational investigations all point towards the fact that morphological parsing strongly depends on the linguistic and distributional properties of the words and their sub-parts. Morphemes are not recognized in isolation but rather relationally in the context of other phonologically similar material. Units in the mind result from contrast, and contrast derives from distributional diversity (Baayen, 2003; Libben and Jarema, 2004; Albright, 2009). The speakers' morphological competence is mainly based on probabilistic reasoning; morpholexical processing is affected to a great extent by the statistical properties of roots and affixes (relative frequency, phonological/orthographic neighborhood density, family size, family frequency, etc.) (Schreuder and Baayen, 1997; Baayen, 2003). Morphemes therefore compete for recognition. The processing of function and lexical elements clearly exemplifies this competition. Simplistically, function words are much more frequent than their nearest lexical neighbors, thus they escape the inhibitory effects of high neighborhood density. Therefore, the processing of function words is predicted to be relatively efficient because of their relative frequency (which is high) and in spite of their lexical neighborhood (which can indifferently be dense or sparse) (Segalowitz and Lane, 2000).

For an inflecting language such as Italian, morpholexical routines based on the distributional diversity of morphological elements have been repeatedly found to be an efficient and frequently activated processing strategy for both word recognition and word naming (Burani and Laudanna, 2003).
The quantitative aspects of morpheme competition also appear to be cognitively relevant in terms of their positional correlates in word structure. Like many other Indo-European languages, particularly of the fusional type, Italian is a predominantly suffixing language; in the majority of the cases, the locus of the expression of morphosyntactic relations is the end of the word. Suffixation is preferred over prefixation in derivation as well. Consequently, functional elements 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 substrings that tend to occur in either position of the word. In other words, one and the same phonological sequence will define a set of different quantitative properties (e.g., absolute "token" frequency, frequency of the lexical forms in which it appears, or number of neighbors) depending on its position in the word (see Table 1 for some examples). These differences will necessarily impact lexical processing as a whole.

ATO#		#ATO	
*nitr**ato***	NITRATE. m. s. "nitrate"	***ato**mico*	(adj.) "atomic"
*mang**iato***	EAT. p. part. "eaten"	***ato**ssico*	(adj.) "non-toxic"
*am**ato***	LOVE. p. part. "loved"	***ato**llo*	(n. sg.) "atoll"
*dat**ato***	GIVE. p. part. "paid" and DATA. m. s. "data"		
*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.*

2. A new concept of formal similarity

In previous experiments, we investigated how native Italian speakers process positional (in addition to quantitative) variables in the decomposition of complex pseudo-words and found that positional regularities at the sublexical level represent psychologically and computationally salient pre-conditions for morphological parsing in that language (Calderone *et al.*, 2008; Calderone and Celata, 2010). The salience of the right edge of morphologically complex pseudo-words (i.e., the portion usually occupied by functional elements) emerged as a by-product of micro-phonotactic preferences and macro-phonotactic positional information. By micro-phonotactics we mean sequential information among segments (e.g., the fact that, in the specific language, a phonological sequence, such as /ato/, differs from similar sequences, such as /uto/, /rto/, and /atu/). By macro-phonotactics we mean positional information within the word, i.e., sub-lexical (or chunk) effects (e.g., the fact that word-initial /#ato/ is different from word-medial /-ato-/, as well as from word-final /ato#/).

These definitions lead to the conclusion that, from the point of view of the language user, the "morphological structure" of words is not the sum of a certain amount of individual, fully represented morphemes concatenated to form a morpho(phono)logically complex word (which in turn pertains to a given morpho-syntactic category and establishes paradigmatic relationships with its category-mates); rather, it derives from a fine-graded representation of phonological similarities and distributional correspondences among a hypothetically open lexical dataset. This position is reminiscent of word-based and output-oriented approaches of morphological analysis (Aronoff, 2007; Blevins, 2006) and is compatible with the idea of a structural continuum spanning from atomic (lexical) to complex (morphological) constructions, that include morphological and possibly larger (syntactic) constructions (Booij, 2010).
In terms of lexical processing, a similar view of morphological complexity implies that, for any given phonotactic sequence that a speaker individuates as a possible access unit to the morphological structure of a word (e.g., word-final /ato#/), a complex network of associations arises, involving not only similar sequences in the same position of similar words (all occurrences of word-final /ato#/ plus, e.g., word-final /uto#/ or word-final /to#/), but also similar sequences in different positions (e.g., word-initial /ato#/). The strength of each association will depend not only on phonological similarity and positional coincidence, but also, crucially, on token frequencies and on a chain of neighborhood effects. In the absence of frequency information, a simple mechanism of segment substitution, insertion, deletion (such as a traditional edit distance algorithm) would produce similarity measures among strings, which would allow the speaker to derive at least general assumptions about the phonotactic preferences of the language. In fact, when provided with frequency information at the *type* level, probabilistic phonotactics has been repeatedly shown to shape the native speakers' reactions to different phonological patterns attested in appropriately constructed lists of non-word stimuli (mostly consisting of monosyllables covering a limited set of phonological options with respect to word-initial and word-final combinations) (Bailey and Hahn, 2001; Frisch and Zawaydeh, 2001; Hayes and Wilson, 2008; Adriaans and Kager, 2010). However, such a mechanism appeals solely to contrastive principles tuning the speaker's reactions to the effects of systematic contrasts in a given language. This sort of "phonotactic contrast" focuses on the space of phonotactic generalizations that a speaker can derive from a process of comparing different *types* of attested sequences of phonemes (versions of these generalizations are algorithmically modeled by the *Minimal Generalization Learner*) (Albright and Hayes, 2003)). Theoretically these approaches assume that the token frequency of phonotactic patterns within the word is normalized, or at least uniform, according to distributional grounds. Still, some authors explicitly reject the idea that token frequency may play a role in phonotactic processing (Albright, 2009).

We claim that phonotactic regularities that are more frequently attested in a representative corpus of a given language shape the speakers' phonotactic knowledge by chunking sequences of phonemes in a fine-grained fashion. Besides the "paradigmatic contrast" that emphasizes the role of type frequency, a sort of "syntagmatic contrast" expressed by different values of token frequency may highlight structural regularities at the interface of phonology and morphology in a given language. If one considers, for example, the minute data set represented in Table 1, the very high frequency of occurrence attested for forms such as *amato* "loved" and *mangiato* "eaten" in any Italian corpus (as well as for their phonologically and paradigmatically related forms, such as *amata* "loved.fem.sing", *mangiata* "eaten.fem.sing", *amati* "loved.m.pl." etc.), compared to the relatively lower frequency of occurrence for forms such as *atomico* "atomic" or *atossico* "non-toxic", provides an unambiguous cue for the different morphological status of final /atV#/ with respect to initial /#atV/. Token frequency information appears, therefore, to contribute in isolating functional units on the basis of their processing saliency inside the word and in probabilistically deriving their ortho-phonological boundaries.

For this reason, our corpus-based psycho-computational account includes token frequencies as a fundamental parameter in shaping the multivariate statistical structure of the environment. A new concept of formal similarity holding among words and sublexical structures is developed, which derives from a non-linear interaction of multiple factors affecting the properties of the phonological string (phonemic content, phonotactic regularities, and 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, and token frequency distributions). This enriched notion of formal similarity, which is in no way a remnant of an edit distance algorithm as a measure for superficial similarities, is modeled in our algorithm in terms of an activation-based lexical representation, dealing with both local/phonotactic (string-level) and global (word-level) constraints (see below, § 3.1).

According to our hypothesis, this complex nucleus of relationships that exists among words turns out to provide the speakers of a language with a surplus of information, which gives rise to proto-morphological representations of lexical forms and provides a cognitively efficient root to complex word decomposition.

As in previous experiments (Calderone *et al.*, 2008; Calderone and Celata, 2010), this hypothesis is tested both behaviorally and computationally, using an experimental protocol aimed at correlating the speakers' responses with the computational output obtained over the same linguistic data set. In particular, morphologically complex pseudo-words are used to elicit ortho-phonological similarity values from both native Italian subjects and the activation-based computational device that was trained on a phonologically encoded corpus of written Italian.

It must be noted that, in spite of its apparent *naiveté*, the word similarity judgment task is a complex cognitive task, with implications that go well beyond the registration of the speakers' unconscious reactions in conditions of induced behavioral pressure (such as those of many online experimental tasks). The word similarity judgment task involves clear linguistic and meta-linguistic processing strategies and simulates the kind of operations a speaker usually performs in natural language processing. Therefore, word similarity elicitation has the two merits of naturalness and robustness, which other experimental procedures do not show to the same extent, at least for the present purposes [1].

Similarly, on the computational side, the system is asked to extract similarity values for vectorial representations of words (see below, § 5, for details). This involves a heavy analytical component over a large-scale input (consisting specifically in a corpus of phonologically transcribed written Italian, including approximately 5 million word tokens) and a generalization process conducted over word-sized units. This procedure is therefore of greater complexity if compared to the general information stripping mechanisms implemented by proposed models of morphological learning, which operate on closed lists of input stimuli and associated semantic features (Rueckl and Raveh, 1999; Plaut and Gonnerman, 2000; Harm and Seidenberg, 2004). As we have already mentioned, the procedure also exceeds the amount of information commonly used in the generalization procedures operated by *n*-gram sampling models of phonotactic learning, which concentrate on clusters or syllables and which abstract away from parameters of word token frequency (Bailey and Hahn, 2001; Hayes and Wilson, 2008; Adriaans and Kager, 2010).

The paper is organized as follows. Section 3 illustrates materials and procedures of the psycho-computational experiment, with a focus on the criteria adopted in the realization of the experimental corpus. Section 4 contains the specific experimental hypotheses and the results of the word similarity judgment test performed by native Italian speakers. In section 5, the computational algorithm is presented: 5.1 reveals the general mathematical formalism implemented in the model (in terms of the learning phase and transfer function), and 5.2 reports specifically on the simulation *scenario* of the present experimentation. Section 6 contains the main results: the computational output is tested against a *Relative Levenshtein Distance* (RLD) baseline and against the native speakers' performance and is analyzed with respect to the experimental hypotheses that were put forth in section 4. Section 7 concludes and proposes future directions of psycho-computational research on morphological processing within an activation-based framework.

[1] A criticism of masked priming procedures, so extensively used for the investigation of morphological effects in word recognition, can be found in (Rueckl, 2010), who points specifically to the extreme sensitivity of masked priming to "subtle methodological variations that seem to have much more to do with general visual processes than with word recognition per se" and argues that, from the point of view of testability, "the slippery slope here is that to get the model's behavior to align with experimental results one would end up 'modeling the task' rather than modeling the word recognition process – an unattractive outcome for many of us who do modeling" (p. 386).

3. A psycho-computational account of morphological decomposition in Italian: experimental materials

3.1. *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. Using a large list of derivational affixes taken from *Gradit* (DeMauro, 1999-2007), 21 prefixes (15 disyllabic and 6 monosyllabic) and 14 suffixes (13 disyllabic and only 1 monosyllabic) were selected. For each of them, the following data were drawn from CoLFIS (Bertinetto *et al.*, 2005): (i) the number of word forms (i.e., lexical contexts) in which the phonological sequence corresponding to the selected affix was present in the initial position; (ii) the number of word forms in which the phonological sequence corresponding to the selected affix was present in the final position; (iii) the 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 the initial position; and (iv) the number of token forms in which the phonological sequence corresponding to the selected affix was present in the final position. Selected examples are given in Table 2. It is worth noting that these quantitative data were calculated with respect to the phonological sequence corresponding to the real affix and not to the affix itself. For example, if the prefix *ambi-* was selected, CoLFIS was asked to provide the number of word forms or tokens beginning with *ambi-* in cases (i) and (iii) (e.g., *ambisillabico* "ambisyllabic" but also *ambizione* "ambition") and the number of word forms or tokens ending with -*ambi* in cases (ii) and (iv) (e.g., *giambi* "iambs"). For this reason, we will use the term "pseudo-affix" (either "pseudo-prefix" or "pseudo-suffix") to refer to phonological sequences such as *ambi* in *ambisillabico*, *ambizione*, *giambi*[2].

The relation between the distributional behavior of pseudo-affixes and the expected behavior of the phonologically coincident real Italian affixes can 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

[2] Stress was not taken into consideration: for example, when considering the number of word forms/tokens with *mono* in final position, both *kimono* (with stress on the second syllable) and *premono* "(they) press" (with stress on the antepenultimate) were taken into account. In fact, given that words in CoLFIS are not phonologically encoded, it became impossible to separate stressed from unstressed phonological sequences. The quantitative data derived from CoLFIS for each relevant sequence have therefore been calculated with respect to prosodically underspecified (i.e., stressless) representations of the relevant word forms.

with 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 of this pattern is the sequence *rico*. *Rico* corresponds to an Italian suffix but, if one looks at the CoLFIS data, this sequence turns out to be much more represented as word-initial than word-final, especially when token forms are considered. This certainly has to do with the fact that the sequence *ri* corresponds to a prefix in Italian, which is quite productive. 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 with the phonological sequence. Finally, the corpus also included pseudo-affixes unevenly attested in terms of word forms vs. token count: for example, the sequence *iso* (which coincides with an Italian prefix) was represented more frequently in word initial than in word final positions when word forms were considered, but the reverse pattern was found when the token count was looked at. We interpreted this latter condition as one of distributional ambiguity (due to a mismatch between word forms and tokens data) coupled with an inconsistency in the relation between the grammatical status of the affix and the rough distributional information associated with the phonological sequence (for token count).

Phonemic sequence	Word forms initial	Word forms final	Token forms initial	Token forms final	Grammatical status	Distributional ambiguity	Consistency distribution/ grammatical status
pre	1,214	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. *Examples of pseudo-affixes and their distributional properties (source: CoLFIS, Bertinetto et al. 2005).*

In conclusion, by considering the quantitative properties of pseudo-affixes in terms of both word forms and tokens and by crossing the two parameters of distributional ambiguity and consistency between the distributional data of pseudo-affixes and the grammatical status of homophonous affixes, we traced the distributional *continuum* back to a grid of interpretable conditions which were used in the analysis presented in this paper (§ 4.2 and § 6 below).

3.2. *Pseudo-words and their association conditions*

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 type-1 associates, the pseudo-affix shared the same position as 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 type-2 associates, the pseudo-affix was split up within the word (e.g., ***p**o**r**n**e**si* was associated to *mula**pre***, and *so**r**n**ep**i* to ***pre**luma*). Type-2 associates were created to evaluate the processing of the two different type-1 associates (where the affix is included in word initial and final positions, respectively) against a neutral baseline and to compare according to that basis the processing of phonotactic similarities in the two different positional conditions.

It is also worth noticing that because each pseudo-affix was placed in both the initial and the final positions of the morphologically complex pseudo-word, the pseudo-affix appeared half of the time in its "licit" position (e.g., ***pre**luma*, ***pre**niso*: "pre" is a prefix in Italian) and the other half in its "illicit" position (*mula**pre***, *soni**pre***). Therefore, pivots and type-1 associates made up of pseudo-prefixes in the initial position can be considered morphological pseudo-words in the typical sense (Burani *et al.*, 1995). When they occur in the final positional option (e.g., *mula**pre***, *soni**pre***), the level of decomposability is undoubtedly inferior (whether it exists at all is exactly the issue here; see § 4.1), 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 and pseudo-morphological pseudo-words when they occur in the initial positional option. Each pseudo-affix was therefore associated with morphological and pseudo-morphological pseudo-words by the same proportion.

As in the case of the pivot items, all associated words in each set were anagrams of each other. The pairs were also balanced with respect to the fact that the initial and final segments of both associates in the non-identity condition (e.g., *so**r**n**ep**i* and ***p**o**r**n**e**si*) were always different from the initial and final segments of the corresponding pivot items.

Stress position was carefully controlled. Stress was always on the penultimate syllable of the pseudo-words. The quality of the stressed vowel was balanced across the identity/non-identity conditions; for example, if the stressed vowel was different in the first associate with respect to the pivot (e.g., ***pre**luma* – ***pre**niso*), it was also different in the second associate with respect to the pivot (e.g., ***pre**luma* – *so**r**n**ep**i*).

			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. *Examples of pseudo-words and the conditions of their associations.*

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 x 2 positional options x 2 identity conditions) experimental pairs was finally created.

4. Behavioral experiment

4.1. *Procedure and analysis*

16 pairs (some of them reproduced the pattern of alternation described above for the experimental pairs, and others were made of two identical non-words, such as *muserota-muserota*) were created and used as fillers. The resulting 164 pairs were randomly arranged 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 pool of 24 native Italian subjects (aged $25 - 35$) was asked to judge the similarity of each pivot item with respect to either of the two associated items.

A beep sound signaled the beginning of the trial. The stimuli were visually and aurally presented. The subjects sat in front of a computer screen and wore headphones, through which they could hear a female native Italian voice pronouncing each pseudo-word pair with an inter-trial interval of $3,500$ msec and an inter-stimulus interval of 100 msec. The trials were also written on a sheet. The subjects were asked to judge the degree of formal similarity between the members of each pair according to a 9-point scale ($1 =$ minimum similarity, $9 =$ maximum similarity or identity) and write the numerical value in a corresponding cell on the answer sheet. The subjects were previously instructed about the range of phonological variation they would encounter in the experiment, and in particular, they were told that the members of each pair did not differ in length and stress position and that some pairs consisted 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 to modulate the use of the numerical values according to the global range of phonological variation in the experiment. Prior to testing, a familiarization phase was added with the specific purpose of allowing individual judgment calibration. The subjects were also informed about the completely offline nature of the experiment, which allowed them to make cross-trial comparisons and even auto-corrections during the experimental session. Finally, they were reassured that their judgments would be absolutely subjective and were 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.

If the perceived similarity of words is the result of a complex nucleus of quantitative relationships among lexical and sublexical structures, such as those referred to above (§ 2), and if this surplus of distributional information allows the speakers to derive at least the general phono-morphological properties of the language, then our Italian subjects should process the final portion of the pseudo-words differently than the initial one, as the former represents the position where functional elements preferentially appear. Thus, phonological variation in word-final position should be judged to have a "higher" processing cost than variation in word-initial position, all other things being equal. In the terms of our experiment, the perceived difference between the identity and the non-identity conditions was expected to be greater for stimuli containing pseudo-affixes in the final position than for those containing pseudo-affixes in the initial position, irrespective of the grammatical nature of the affix. For example, both pairs *preluma-preniso* and *mulapre-sonipre* should elicit higher similarity values than, respectively, *preluma-sornepi* and *mulapre-pornesi*, but the difference between *mulapre-sonipre* and *mulapre-pornesi* was expected to be greater than the difference between *preluma-preniso* and *preluma-sornepi*. In statistical terms, we expected a significant interaction between the two independent variables of identity condition and positional option for both pseudo-prefixed and pseudo-suffixed stimuli.

4.2. *Results*

The similarity ratings given by the subjects were treated as dependent variables and evaluated through univariate ANOVAs 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 similarity 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 unevenly perceived by the subjects, depending on the relative position of pseudo-affixes in the pivots. In particular, as the average values 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). The same effect held strong independently of the grammatical status of pseudo-affixes (pseudo-suffixes or pseudo-prefixes), as shown by the non-significant interaction position*identity*affix ($F(7, 3079) = 0.110, p > .05$).

Recall that the experimental pairs were perfectly balanced with respect to phonological variables, including stress position (§ 3.1). 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 across positional options. 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 the segmental constituency of the stimuli. The present results 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 the final position, it is noteworthy 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.

In conclusion, 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 significant potential for lexical processing. If this hypothesis is correct, the distributional properties of phonological strings (also referred to above as micro- and macro-phonotactic regularities) may act as preconditions for morphological parsing to the extent that they allow the differential salience of the right edge of the word to emerge in the phonomorphological competence of native speakers of Italian.

5. A phonotactic activation system for ortho-phonological word processing (PHACTS)

The computational simulation was grounded in PHACTS (for PHonotactic ACTivation System), a computational model used to establish how ortho-phonological words are learned, processed and retrieved in the mental lexicon of the speakers (Herreros and Calderone, 2007; Calderone *et al.*, 2008). The same set of morphologically complex pseudo-words described in § 3.2 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).

PHACTS is based on the principles of a Self-Organizing Map (SOM), which is a neurocomputing algorithm for multivariate analysis (Kohonen, 2000) where linguistic structures are conceived of as vector/matrix structures in a multi-dimensional space. 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 degree of well formedness of unseen stimuli and even attributing salience to a specific word subpart, disregarding the rest.

Section 5.1 contains the general mathematical framework of the model, following a bipartite structure consisting of: a) the learning algorithm, and b) the transfer function for the activation-based representation of the original vector input. Section 5.2 contains the details of the present simulation, including: a) the development of an Italian phonotactic knowledge, and b) the derivation of ortho-phonological word representations.

5.1. *PHACTS: the learning algorithm*

The physical structure of PHACTS is defined by a set S (with a finite cardinality) of neurons n_{jk} with $1 \leq j \leq J$ and $1 \leq k \leq K$ arranged in a bi-dimensional grid of $S = \{n_{11}, n_{12}, \ldots n\}$, $\|S\| = J$xK. Each neuron n in the grid corresponds to a vector u (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 assume random values and during the learning phase they change these values in order to adapt themselves to the input data.
The PHACTS algorithm can be described by two phases:

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:

$$BMU((i)t) = \arg\max_{jk}(\mathbf{i}(t) \cdot \mathbf{u}_{jk}) \qquad [1]$$

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 the topological adaptation equation given in [2]:

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

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 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:

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

The learning rate parameter controls for the elasticity of the network, and the neighborhood function roughly controls for the area around each best matching unit 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 $t \to 0$.
PHACTS performs a vector mapping of the input data space to the output space defined by the prototype vectors u_{jk} and its position in the bi-dimensional grid S.
After the learning phase, 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 neurons tend to be more highly activated by more frequent (in terms of token frequency) input stimuli during the learning phase. This frequency effect is reflected in the transfer function phase as well, as outlined below.

b) the 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 for unseen stimuli can be derived with respect to the output space previously obtained. This phase implements a linear threshold 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), u_{jk} be the prototype vector of the neuron n_{jk}, and d be a threshold value working as a filter and dampening the frequency gap of highly scored subsequences in the (raw) input (see below).

Let Φ_{jk} be the transfer function of the neuron n_{jk} defined as in [4]:

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

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

Consequently, the activation-based representation of input x defines a sort of distributed representation of the input x, which reflects both the position in the bi–dimensional topological organization of the grid and the activation of the output space of the neurons.

5.2. *Present simulation*

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

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 contains nearly 80,000 word types and 5 millions word tokens. The corpus was phonologically transcribed following a grid of features defining the phonological grammar of Italian. Consonants were specified for place, manner of articulation and voicing while vowels were specified for roundedness, height and anteriority, plus a specific feature referring to their prominence in the word ([± stress]). Stress notoriously is a property of (phonological) words, not just of syllables or, even less so, individual vowels (Lehiste, 1970). For this reason we believe that adding a binary feature to vocalic segments in order to encode the prosodic prominence of a syllable within the word is a simplistic procedure. How to provide the input data with a better representation of lexical stress is therefore open to future research and the present experiment should be considered as explorative with respect to the possibility of encoding suprasegmentals in the phonological equipment of the system.

In our simulation, neurons were trained on n-grams of phonological words. The system produced adjacent neurons for detected phonotactic regularities, i.e., for frequently attested co-occurrences of phonemes specific to the language. The neurons of the map thus developed a topographic profile of language-specific phonotactics on a distributional basis, whereby the most frequent phonotactic patterns were clearly distinguished by the neurons and quantitatively defined by an activation level, which depended on the token-frequency of each relevant phonotactic pattern. N-gram activations were thought of as units in a network of mutually contrastive relations and the representation of each unit depended on the presence of the other units. The phonotactic knowledge developed from the accumulation of individual patterns acted therefore as a sort of collective representation regulating the interaction of "conspecifics" in terms of their "cooperation" and "competition". In fact, the mapping function operated by the system fit the space of the bi-dimensional grid for the cumulative representations of n-grams incrementally.

In this experiment, the size of the bi-dimensional grid was 25 X 35 neurons, and thus $S = 875$. String sampling was fixed at $n = 6$ (hexagrams). Setting a sampling window is actually a matter of empirical preference, and the choice is justified according to the need to find 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 conditions, such as the type and length of the stimuli, the phonotactics of the language under investigation, and the dimension of the map and the number of neurons involved. The common practice is to use bi- or trigram sampling windows to recover phonotactic local constraints in the input data. When a larger sampling window is used, the neurons focus on larger units, thereby recovering as more salient regularities involving a greater number of phonemes than when bi- or trigram windows are used. Furthermore, regularities involving few segments will produce even smaller effects because 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 believed that a 6-gram sampling window could 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 could be taken into in account. However, a 3-gram based simulation was also run for comparative purposes.

Each word of the corpus therefore passed a 6-gram (or 3-gram) sampling and in this form was given in input to the system. Following the *n*-gram sampling, the model develops a topological organization of the stimuli, which reflects a sort of phonotactic knowledge enriched with quantitative information on token frequencies: similar *n*-gram tokens are mapped adjacent to one 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) activation-based lexical representations (transfer function)

On the basis of the topological organization obtained in the learning phase, each neuron of the bi-dimensional grid acted as a 'transfer function' for unseen input items, i.e., for the pseudo-words described in § 3.2.

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 [5]:

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

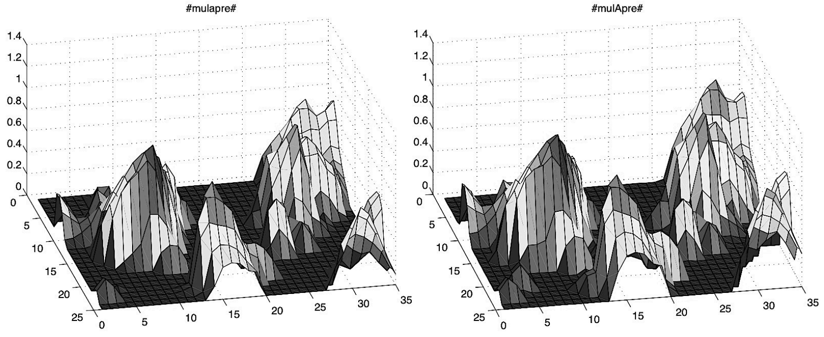

Figure 1. *Comparison between the activation-based representations of the pseudo-word #mulapre#, with no stress specification (on the left) and #mulApre#, with [+ stressed] specification for the second vowel (on the right). In spite of the similarity of the overall profile, slopes and valleys slightly differ as a consequence of the presence/absence of stress codification.*

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 that the number of neurons in the map was $S = 875$), as represented in Fig. 1.

On the basis of such granular and enriched representations, the pseudo-words in the experimental pairs were compared to derive a measure of their formal similarity. This was achieved by adopting the cosine value of the angle between two vector representations, as shown in Fig. 2. Consequently, the similarity scale ranged from 1 (total identity) to 0 (total dissimilarity).

PHACTS's output was evaluated against a simple edit-distance algorithm for string similarity calculation and against the speakers' similarity judgments elicited under the same experimental conditions (§ 4.2).

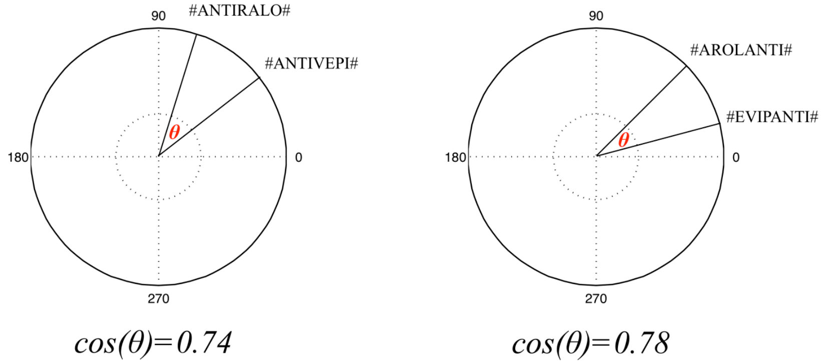

Figure 2. *The cosine value between pairs of pseudo-words is adopted as a measure for pseudo-word similarity.*

6. Human/machine correlation

The speakers' judgments were correlated with the system's output (an inverse log transformed cosine values obtained by the 6-gram sampling) in order to obtain a measure of the psychological reliability of the simulation. In the present experiment, we derived 7 different computational models by varying the activation threshold d from 2 to 5 (in steps of 0.5). The present results refer to the 4-thresholded version of the model, which performed the best with respect to the correlation score with the speakers' performance. As already mentioned, the role of the threshold (d) is that of dampening the differences in frequency among phonotactic sequences. Because it is a corpus-based model, PHACTS is fed by raw and unbalanced phonological words of a corpus and exploits the phonotactic regularities of the language for deriving a lexical phonological representation of each word. The authors purposely have left the corpus unbalanced; using a corpus that is, at least in principle, representative of the use of the Italian language emphasizes indeed the robustness of the PHACTS algorithm. The magnitude of word frequency in an unbalanced raw corpus can bias the output of the computational system and yield "saturation effects" that occur when frequencies are excessively high for some specific words (and for the relative phonotactic sequences). For this reason, the threshold value (d) has been implemented to take control of the distributional variance concerning the phonotactic patterns.

The obtained Spearman's *rho* coefficients (and statistical power) are summarized in Table 4.

A strongly significant correlation was found overall $(\rho = .605, p < .001)$ and the correlation coefficients were even higher in the subsets of items made of distributionally unambiguous pseudo-prefixes $(\rho = .703, p < .001)$ and grammatically/distributionally consistent pseudo-prefixes $(\rho = .621, p < 001)$. These circumstances indicate that PHACTS was strongly guided by the statistical cues present in the training corpus; when such cues contained elements of ambiguity or inconsistency, PHACTS's generalization abilities appeared to be undermined and the system's performance deviated more from the speakers' performance.

PHACTS correlated with the speakers' response to a larger extent than the baseline condition represented by a *Relative Levenshtein Similarity* (RLD). The Levenshtein distance between two strings o and d is defined as the number of insertions, substitutions or deletions necessary to convert the string o into the string d. In our simulations, we adopted a *Relative Levenshtein Similarity* that is equal to $1-\mathrm{RLD}(o, d)$, corresponding to the classical Levenshtein distance (LD) normalized by the sum of the number of characters in both strings, that is, $\mathrm{LD}(o, d)/(|o| + |d|)$.

Recall that in our experiment, type-1 and type-2 associates were anagrams of one another; both consisted of the same segmental material disposed in different sequential order. RLD was used to calculate the formal similarity between the *pivot* items and each of their associates, taking into account the variations in segments' sequential order. RLD turned out to fit the speakers' response to a reduced extent $(\rho = .464, p < .001)$. On the contrary, PHACTS appeared to approximate the speakers' behavior with a higher level of accuracy, outperforming RLD and showing that the activation-based lexical representations generate a surplus of information directly exploitable for complex word decomposition, reflecting a process similar to that demonstrated by native speakers.

Two control simulations were run, with the purpose of establishing the role of two independent parameters on the generalization abilities exhibited by PHACTS. The first parameter had to do with the mechanical procedure of input data processing, with particular reference to the size of the *n*-grams window used by the system to sample the input strings. The second one referred to the type of phonological information that the system exploited to process the data; in particular, we wanted to verify whether and to what extent lexical stress codification did help the system to generate reliable lexical representations for the input data.

The size of the sampling window was varied, with the purpose of determining the difference between our 6-gram and the classical 3-gram sampling procedure. A simulation was run that was identical to the one described above (except for the size of the sampling window). As shown in Table 4, the correlation coefficient dramatically dropped off with respect to the simulation based on 6-grams $(\rho = .227, p < .001)$.

This result corroborated our view that using large sampling windows for richly specified phonological representations allowed the system to capture crucial long-distance relations among segments and segmental chunks.

In addition, in order to ascertain the role of stress codification in the phonological representation of input words, a simulation was run that was exactly identical to the one described above (6-gram sampling window, 4-thresholded version), except that it did not include the [± stress] feature in the phonological codification of input vowels. In this simulation, therefore, both the training corpus and the experimental materials were phonologically encoded in such a way that vowels were only specified for roundedness, height and anteriority. As indicated in Table 4, the correlation coefficient was lower when compared to the one obtained when the input data were specified for stress $(\rho = .504, p < .001)$. This result demonstrated that introducing a [± stressed] feature in the phonological codification of input vowels, though being an unsophisticated procedure to encode lexical stress, improved the performance of the system with respect to the speakers' performance.

		N	**Spearman's *rho***	**p**
RLD*speakers correlation		3,080	.464	.000
PHACTS*speakers correlation	With stress codification (6-grams)	3,080	.605	.000
	Without stress codification (6-grams)	3,080	.504	.000
	3-gram sampling with stress codification	3,080	.227	.000
Distributionally unambiguous		2,728	.703	.000
Distributionally ambiguous		352	.272	.021
Distrib./gramm. consistent		2,200	.621	.000
Distrib./gramm. inconsistent		880	.452	.005

Table 4. *Spearman's rho correlation coefficients between the speakers' similarity judgments and Relative Levenshtein Similarity (RLD) values (on the top) and between judgments and the PHACTS inverse log transformed cosine values (on the bottom). Unless otherwise specified, the system's output derives from a 6-gram sampling with stress specification.*

To the extent that the system diverged in a non-negligible way from natural language processing (as implied by the .605 overall correlation coefficient), more factors deserve further investigation.

Figure 3 illustrates the median, standard deviation and range of the speakers' judgments and of the PHACTS's similarity values in the four relevant conditions. Splitting the data for the two experimental factors (identity condition and positional option) revealed that the system was less accurate at generalizing similarity values in the non-identity condition (both final and initial positions). In that case, the cosine values covered an area of relatively greater dispersion, compared to the speakers' judgments, indicating that the system exhibited more fluctuations in assigning similarity values to pairs of non-words sharing the same phonemic sequence in different internal positions, while the speakers were more uniform in assigning low similarity values to non-identical pairs (see Figure 3).

For these reasons, the correlation coefficients for the four subgroups of data are very different from one another, as visualized in Figure 4.

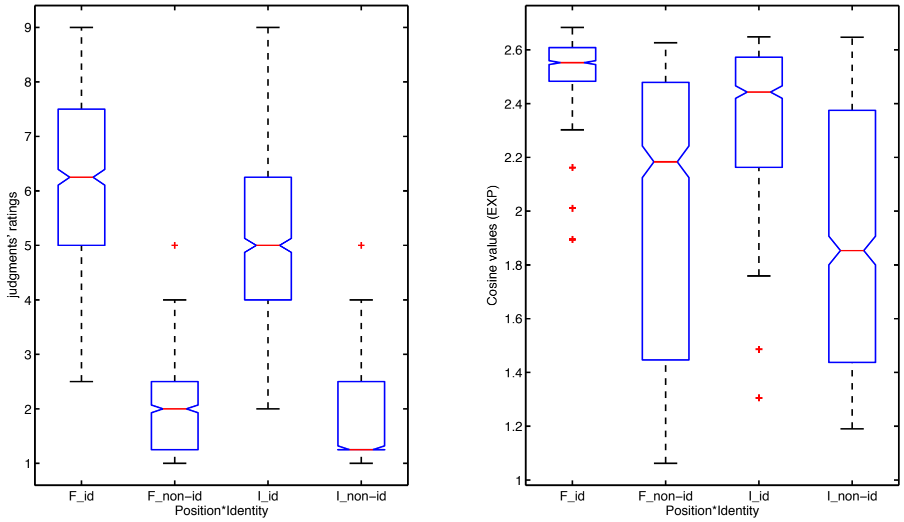

Figure 3. *Median, standard deviation and range values for the speakers' judgments (left box) and the exp-transformed cosine values (right box), split by position and identity.*

An affix-by-affix analysis also showed that there were areas of unpredictability in the experimental set. The values in Table 5 show that for most affixes PHACTS correlated with the speakers' judgments on a higher level than RLD distance did. Yet this trend could not be generalized to the entire data set, and there were also cases where PHACTS produced unreliable similarity values, as demonstrated by non-significant or negative correlation values. In order for PHACTS to reach higher levels of accuracy, these sources of unpredictability in the system's functioning will require closer investigation.

Finally, in order to ascertain whether PHACTS was sensitive to the differential salience of the right edge of the word to the extent that native speakers of Italian resulted to be (see § 4.2), an analysis of variance was run with the exp-transformed cosine values (6-gram sampling; activation threshold 4) as the independent variable. As expected, there was a significant main effect of the identity factor (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). In conclusion,

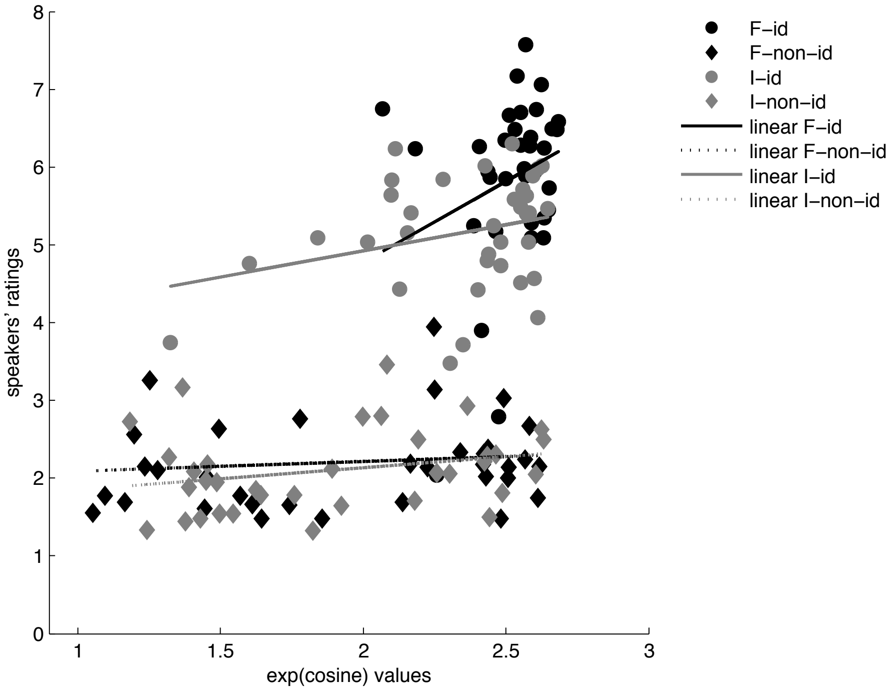

Figure 4. *Speaker/system correlation split by position and identity.*

PHACTS's sensitivity to the differential salience of the right edge of the word was similar to that shown by the native speakers, but of an inferior magnitude.

ambi	**ana**	**anti**	**arci**	**bis**	**rino**	**tino**
RLD .496**	RLD .510**	RLD .504**	RLD .812**	RLD .746**	RLD .668**	RLD .600**
PHACTS .730**	PHACTS .379**	PHACTS .667**	PHACTS .636**	PHACTS .95**	PHACTS .746**	PHACTS .641**
cata	**co**	**ello**	**emi**	**endo**	**tico**	**to**
RLD .547**	RLD .431**	RLD .850**	RLD .822**	RLD .805**	RLD .760**	RLD .276**
PHACTS .784**	PHACTS .502**	PHACTS .820**	PHACTS .017**	PHACTS .780**	PHACTS -.185**	PHACTS .348**
ento	**epi**	**filo**	**ico**	**in**	**tore**	**rico**
RLD .783**	RLD .756**	RLD 756**	RLD .809**	RLD -.241**	RLD .738**	RLD .602**
PHACTS .800**	PHACTS .731**	PHACTS .771**	PHACTS .746**	PHACTS .210**	PHACTS .661**	PHACTS .686**
ino	**intra**	**ione**	**ipo**	**ismo**	**ri**	**pro**
RLD .717**	RLD .713**	RLD .796**	RLD .839**	RLD .831**	RLD .476**	RLD .720**
PHACTS .653**	PHACTS .724**	PHACTS .552**	PHACTS .445**	PHACTS .754**	PHACTS .568**	PHACTS .650**
iso	**mega**	**mento**	**one**	**ore**	**pre**	**peri**
RLD .597**	RLD .818**	RLD .854**	RLD .777**	RLD .793**	RLD .759**	RLD .525**
PHACTS .616**	PHACTS .760**	PHACTS .719**	PHACTS .494**	PHACTS .444**	PHACTS .777**	PHACTS .589**

Table 5. *Spearman's rho coefficients for speakers-RLD and speakers-PHACTS (exp-transformed cosine values) correlations, split by pseudo-affix.*

Given the numerical dispersion associated with the variability in the system's output, we wanted to explore the possibility that the expected identity*position interaction did hold true for some empirically defined subset of the data. Our first hypothesis was that the system was likely to be more sensitive to the distributional information of the input than the native speakers. We therefore wanted to check whether different types of affixes elicited different responses from the system. The two parameters of distributional-grammatical consistency and of distributional ambiguity were thus used as possible predictors of the system's reactions.

The first parameter was found not to influence the response pattern because 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 because 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). We can consequently conclude that, when there was no mismatch between different sources of quantitative information relative to the training data, PHACTS was able to let the salience of the right edge of the word emerge, corresponding with the consistent practice of native speakers of Italian.

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, was therefore considered to be a further predictor of the system's reactions to the similarity evaluation task. The length factor was thus included in the analysis to establish 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. However, the data refuted this hypothesis: 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. This indicated that the length of the pseudo-affix was irrelevant for the transfer function operated by PHACTS on novel stimuli, at least as far as the final position salience effect was concerned.

7. Conclusions

In this paper we have investigated the role of multiple quantitative properties of phonological strings for the decomposition of morphologically complex words. The problems we have addressed have to do with the so-called morphology-phonology interface as well as with the lexical bases of morphological contrast. Thus, to paraphrase the issue title, all these problems are "toward" and "beyond" morphology at the same time. With respect to the morphology-phonology interface, there is a point where phonotactic pressures and whole-word connections converge and their cumulated effects allow the generalization that the word in question contains functional units. We consider this as a precondition for the development of a morphological competence. On the other hand, with respect to the lexical bases of morphological contrast, the hypothesized activation effect is theoretically antecedent and independent from form-to-meaning morphological mapping in the classical sense.

Italian was taken as an example of inflectional language, and proto-morphological routes of word decomposition, accounting for the emergence of the right edge of the word as a psycho-computationally salient access unit based solely on micro- and macro-phonotactic effects, were found to be effective in lexical processing by humans as well as in word-level generalizations operated by PHACTS.
PHACTS modelization demonstrated that proto-morphological meaning may emerge from a cooperation of local/phonotactic (string-level) and global (word-level) constraints, uncovering the dynamics of lexical and sub-lexical connections within a multivariate domain. Future analyses should be directed to non-inflecting languages (or to languages in which morphological elements are not predominantly suffixes) in order to test the role of micro- and macro-phonotactic profiles in different conditions of morphological complexity. The psychological plausibility of PHACTS was established by directly comparing the output of the system's generalization with the native speakers' word similarity ratings. Importantly, the same set of experimental stimuli was used for the behavioral and the computational testing and the experimental procedure was kept as uniform as possible in the two tasks. To the best of our knowledge, no comparable efforts can be found in the psycho-computational literature in terms of complex cognitive tasks. The only exceptions may be seen in the framework of phonotactic modeling and wordlikeness studies (Scholes, 1966), where some recent attempts have been made at correlating the speaker with the machine in processing lists of monosyllabic words (Hayes and Wilson, 2008) or strings of unsegmented input (Adriaans and Kager, 2010). Our computational model has the merit of allowing corpus-based, instead of list-based, simulations.

The consequences of recovering the phono-morphological profile of a language should also be tested with respect to gradient wordlikeness effects, such as those revealed by a conspicuous number of phonotactic modeling studies. The ability of PHACTS to produce wordlikeness comparative judgments for monosyllabic English non-words is currently under investigation.
Several empirical questions raised by PHACTS's structure and functioning also call for further investigation. These concern principally the treatment of the statistical distributions that shape the input data and its effects on the lexical representation of the word. Token frequencies, in particular, need to be carefully calibrated in the model, possibly according to a non-linear function, to smooth the results of highly dispersed values. Furthermore, the dampening effect operated by the currently implemented threshold function will need further refinement.

Finally, we believe that the present results have implications for both theories of lexical access and output-oriented 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 isolation

of the constituting elements of a the word. A "suffixation preference" (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 *et al.*, 2010).
With respect to probabilistic and output-oriented models of morphology, this study supports the idea that distributional effects, particularly those related to positional variations at the sub-lexical level, should be accounted for in (interactive, hybrid) models of morpholexical processing. Interactions between "routes" or "levels" are organized with respect to string-specific characteristics extending beyond the effects of well-known quantitative properties of individual morphemes, such as their relative frequency, neighborhood density, and family size, and involving a graded notion of morphological salience of the access units as an additional (possibly confounding) source of processing information.

Acknowledgements

We are grateful to Pier Marco Bertinetto, Wolfgang "Ulli" Dressler, Alessandro Laudanna and four anonymous reviewers for their comments on an earlier version of this article.

8. References

Adriaans F., Kager R., “Adding generalization to statistical learning: The induction of phonotactics from continuous speech”, *Journal of Memory and Language*, vol. 62, p. 311-331, 2010.

Albright A., “Modeling analogy as probabilistic grammar”, *in* J. P. Blevins, J. Blevins (eds), *Analogy in Grammar: Form and Acquisition*, Oxford University Press, Oxford, 2009.

Albright A., Hayes B., “Rules vs. analogy in English past tenses: A computational/experimental study”, *Cognition*, vol. 90, p. 119-161, 2003.

Aronoff M., “In the beginning was the word”, *Language*, vol. 83, no. 4, p. 803-830, 2007.

Baayen H. R., “Probabilistic approaches to morphology”, *in* R. Bod, J. Hay, S. Jannedy (eds), *Probabilistic Linguistics*, MIT Press, Cambridge MA, p. 229-287, 2003.

Bailey T. M., Hahn U., “Determinants of wordlikeness: Phonotactics or lexical neighborhood?”, *Journal of Memory and Language*, vol. 44, p. 568-591, 2001.

Bertinetto P. M., Burani C., Laudanna A., Marconi L., Ratti D., Rolando C., Thornton A. M., “Corpus e Lessico di Frequenza dell’Italiano Scritto”, 2005.

Blevins J. P., “Word-based morphology”, *Journal of Linguistics*, vol. 42, p. 531-573, 2006.

Booij G., *Construction Morphology*, Oxford University Press, 2010.

Burani C., Laudanna A., “Morpheme-based lexical reading: Evidence from pseudo-word naming”, *in* E. Assink, D. Sandra (eds), *Reading Complex Words: Cross-Language Studies*, Kluwer, Dordrecht, p. 241-264, 2003.

Burani C., Thornton A. M., Iacobini C., Laudanna A., “Investigating morphological nonwords”, *in* W. Dressler, C. Burani (eds), *Cross-Disciplinary Approaches to Morphology*, Kluwer, Dordrecht, p. 241-264, 1995.

Calderone B., Celata C., “The morphological impact of micro- and macro-phonotactics. Computational and behavioral analysis”, *14th International Morphology Meeting*, Budapest, May, 2010.

Calderone B., Celata C., Herreros I., “Recovering morphology from local phonotactic constraints”, *Laboratory Phonology 11th Conference "Phonetic Details in the Lexicon"*, Wellington, New Zealand, June-July, 2008.

Caramazza A., Laudanna A., Romani C., “Lexical access and inflectional morphology”, *Cognition*, vol. 28, no. 3, p. 297-332, 1988.

Crepaldi D., Rastle K., Davis C. J., “Morphemes in their place: Evidence for position specific identification of suffixes”, *Memory and Cognition*, vol. 38, no. 3, p. 312-321, 2010.

DeMauro T. (ed.), *Grande Dizionario Italiano dell’Uso (GRADIT)*, Einaudi, 1999-2007.

Frisch S., Zawaydeh B., “The psychological reality of OCP-Place in Arabic”, *Language*, vol. 77, p. 91-106, 2001.

Giraudo H., Grainger J., “A supralexical model for French derivational morphology”, *in* E. Assink, D. Sandra (eds), *Reading Complex Words: Cross-Language Studies*, Kluwer, Dordrecht, 2003.

Harm M. W., Seidenberg M. S., “Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes”, *Psychological Review*, vol. 111, no. 3, p. 662-770, 2004.

Hayes B., Wilson C., “A maximum entropy model of phonotactics and phonotactic learning”, *Linguistic Inquiry*, vol. 39(3), p. 379-440, 2008.

Herreros I., Calderone B., “Spatial lexical representations and unsupervised bootstrapping in morphology”, *Workshop on Machine Learning and Cognitive Science of Acquisition*, University College, London, June 21-22, 2007.

Hupp J., Sloutsky V., Culicover P. W., “Evidence for a domain general mechanism underlying the suffixation preference in language”, *Language and Cognitive Processes*, vol. 24, p. 876-909, 2009.

Kohonen T., *Self-Organizing Maps*, Springer, Heidelberg, 2000.

Lehiste I., *Suprasegmentals*, MIT Press, 1970.

Libben G., Jarema G., “Conceptions and questions concerning morphological processing”, *Brain and Language*, vol. 90, no. 1–3, p. 2-8, 2004.

Meunier F., Segui J., “Cross-modal morphological priming in French”, *Brain and Language*, vol. 81, no. 1-3, p. 89-102, 2002.

Plaut D. C., Gonnerman L. M., “Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing?”, *Language and Cognitive Processes*, vol. 15, p. 445-485, 2000.

Quasthoff U., Richter M., Biemann C., “Corpus portal for search in monolingual corpora”, *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, 2006.

Rastle K., Davis M. H., “Morphological decomposition based on the analysis of orthography”, *Language and Cognitive Processes*, vol. 23, no. 7-8, p. 942-971, 2004.

Rueckl J. G., “Connectionism and the role of morphology in visual word recognition”, *The Mental Lexicon*, vol. 5(3), p. 371-400, 2010.

Rueckl J. G., Raveh M., “The influence of morphological regularities on the dynamics of a connectionist network”, *Brain and Language*, vol. 68, no. 1-2, p. 110-117, 1999.

Rumelhart D. E., McClelland J. L., “On learning the past tense of English verbs”, *in* J. McClelland, D. Rumelhart, T. P. R. Group (eds), *Parallel Distributed Processing: Explorations in the Microstructures of Cognition. Psychological and Biological Models*, vol. 2, MIT Press, Cambridge MA, 1986.

Scholes R. J., *Phonotactic Grammaticality*, Mouton, 1966.

Schreuder R., Baayen H. R., “Modeling morphological processing”, *in* Feldman (ed.), *Morphological Aspects of Language Processing*, Lawrence Erlbaum, Hillsdale, 1995.

Schreuder R., Baayen R. H., “How complex simplex words can be”, *Journal of Memory and Language*, vol. 37, no. 1, p. 118-139, 1997.

Segalowitz S. J., Lane K. C., “Lexical access of function versus content words”, *Brain and Language*, vol. 75, no. 3, p. 376-389, 2000.

Taft M., “Lexical access via an orthographic code: The basic orthographic syllable structure”, *Journal of Verbal Learning and Verbal Behavior*, vol. 18, no. 1, p. 21-39, 1979.

Taft M., “Interactive-activation as a framework for understanding morphological processing”, *Language and Cognitive Processes*, vol. 9, p. 271-294, 1994.

A. Materials of the experiments

Phonemic sequence	**Word forms Initial**	**Word forms Final**	**Tokens Initial**	**Tokens Final**	**Grammatical status**	**Pivot Final**	**Associate 1 Final**	**Associate 2 Final**	**Pivot Initial**	**Associate 1 Initial**	**Associate 2 Initial**
ambi	50	8	1,250	408	prefix	arolambi	evipambi	ibimpave	ambiralo	ambivepi	ivambèpi
ana	115	535	920	5,350	prefix	emodana	isutana	asunati	anamedo	anatusi	isanuta
anti	368	525	1,840	6,300	prefix	arolànti	evipànti	iventàpi	antiràlo	antivèpi	avintèpi
arci	27	223	108	669	prefix	imodarci	esutarci	irtusace	arcidomi	arcituse	irtusace
cata	68	199	476	1,393	prefix	damocata	tisecata	chesitati	catadamo	cataseti	chesatita
co	5,223	1,461	125,352	24,837	prefix	fanedanco	rilutisco	roscutili	codafenna	costurili	roscutili
emi	60	82	540	1,722	prefix	arilemi	esutemi	imusete	emirali	emisute	itesume
endo	15	467	225	4,670	prefix	pisomendo	lafurendo	dunforela	endomospi	endofurla	duleranfo
ento	5	691	8	15,893	prefix	pisomento	lafurento	tunforela	entomospi	entofurla	tuleranfo
epi	65	5	520	30	prefix	mirulepi	tasovepi	vipaseto	epimurli	epivosta	vopaseti
filo	50	15	450	105	prefix	catofilo	debifilo	libodife	filocato	filodebi	libodife
intra	85	0	340	0	prefix	fanodintra	supelintra	parlusinte	intrafonda	intrasulpe	perlusinte
in	4,665	334	51,315	2,338	prefix	rupaldin	tobosvin	bivosten	indulpa	invosbe	vibosne
ipo	54	19	648	931	prefix	mirulipo	tasovipo	potosiva	ipomurli	ipovosta	opasvito
iso	57	40	684	2,860	prefix	miruliso	tanoviso	vosonita	isomurli	isovonta	vosonita
mega	35	3	70	9	prefix	teromega	fusimega	gamusefi	megatero	mefasufi	gamusife
peri	65	16	1,755	112	prefix	litoperi	manuperi	litoperi	perilito	perimuna	rupameni
ri	4,504	1,946	45,040	35,028	prefix	mistilari	fenduneri	indurfene	ristilami	rindufene	indurfene
pre	1,214	12	20,638	4,752	prefix	mulapre	sonipre	pornesi	preluma	preniso	sornepi
pro	1,250	11	27,500	132	prefix	mulapro	sonipro	pornosi	proluma	proniso	sornopi
bis	92	3	2,024	9	prefix	munobis	terabis	bersati	bismuno	bistera	tersabi
tino	0	0	245	1,715	suffix	damotino	risetino	retonisi	tinodamo	tinorise	retonisi
rino	5	20	177	1,416	suffix	damorino	tiserino	esintiro	rinodamo	rinotise	esintiro
ino	67	670	1,241	12,410	suffix	tranogino	sberutino	ronustibe	inotrangio	inortusbe	ronustibe
ello	0	0	298	9,536	suffix	tranogello	sbirutello	loroltesbi	ellongiotra	ellosburti	loriltesbo
tico	0	0	387	3,483	suffix	veretico	samutico	camusito	ticovere	ticomusa	tomusica
rico	358	3,938	102	1,428	suffix	valerico	mosurico	comusiro	ricovale	ricomuso	rosimuco
ico	8	24	989	12,857	suffix	tranogico	sberutico	curostibe	icotrangio	icortusbe	curostibe
ismo	0	0	458	2,290	suffix	fogarismo	teremismo	mestomire	ismorgafo	ismorteme	mistomere
ione	1	3	1,556	37,344	suffix	maralione	porisione	porienosi	ionemarla	ionersipo	porieniso
one	20	220	2,291	43,529	suffix	varalone	pobisone	pobenosi	onevarla	onesbipo	esbonipo
tore	3	9	574	7,462	suffix	velitore	samutore	sumerota	torelevi	toremusa	someruta
ore	30	300	778	16,338	suffix	velimore	tanudore	rudetona	orelvemi	orenduta	etondura
mento	6	12	588	11,760	suffix	rupamento	delimento	mindolete	mentorupa	mentolide	tendolemi
to	690	7,590	4,827	135,156	suffix	festinito	marcureto	morcutera	tonistife	torcurema	morcutera

Table 6. *Pseudo-affixes and their distributional properties in the Italian lexicon. Source: (Bertinetto* et al.*, 2005). Rightmost columns: pseudo-words used in the behavioral and computational experiments.*