SCUOLA
NORMALE
SUPERIORE
PISA

Basilio Calderone & Chiara Celata

The morphological impact of micro- and macro-phonotactics

Computational and behavioral analysis

(talk given at 14th International Morphology Meeting*, Budapest, May, 13-16, 2010)*

0 Outline

This paper investigates the morphological impact of the quantitative properties of the lexicon in the decomposition of morphologically complex words by native speakers of Italian. It deals with a definition of the crucial notions of micro- and macro-phonotactics with respect to word structure, and illustrates the results of a word similarity judgment experiment, where the native speakers' performance on morphologically complex pseudo-words was compared to the output of an activation-based model trained on the same experimental material. We discuss the hypothesis that morpholexical processing is based to a great extent on statistical preconditions that are intrinsic to the micro- and macro-phonotactics of the language.[*]

1 Introduction: On morpholexical processing

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). Early affix-stripping mechanisms account for the non-semantically-driven component of the morpheme recognition process, whereby units in the mind result from *contrast*, and contrast derives from distributional diversity (e.g., Baayen 2003, Libben & Jarema 2004). In this view, morpholexical processing is to a great extent affected by the statistic properties of the lexicon (or sub-parts of the lexicon), and primarily by quantitative

[*] Thanks to W.U. Dressler and H. Howard for useful comments on the occasion of the oral presentation at *IMM14*, Budapest, 2010.

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.). The quantitative properties of the affixes may even interact with later semantic effects, such as those related to family transparency (e.g., Boudelaa & Marslen-Wilson 2009).

Morphemes, then, *compete* for recognition: for example, function and lexical morphemes compete with each other for recognition, but since function words are much more frequent than their nearest lexical neighbours, they escape the inhibitory effects of high neighborhood density. Therefore, the processing of function words is predicted to be relatively efficient thanks to relative frequency (which is high) and in spite of lexical neighborhood (which can be high or low regardless) (Segalowitz & Lane 2000).

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 way processing strategy for both word recognition and – more recently – word naming (see Burani & Laudanna 2003 for a recent review).

In this contribution, we would like to further investigate the quantitative aspects of morpheme competition in terms of their *positional correlates* in Italian word structure. 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 be trivially 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 neighbours etc.) depending on its position in the word (Table 1). These differences will necessarily impact over lexical processing altogether.

Given these premises, the paper addresses the two following research questions: How are positional variables (beside quantitative ones) processed in decomposing complex words? Do they represent psychologically and computationally salient *pre-conditions* for morphological parsing in Italian? By answering these questions, we believe we will be able to empirically test some aspects of the emergent nature of morpholexical processing of complex words in natural (inflecting) languages (Bybee 2007; McClelland et al. 2002).

ATO #		ATO#	
mangi**ato**	EAT.p.part. ‘eaten’	**ato**mico	(adj.) ‘atomic’
amato	LOVE.p.part. ‘loved’	**ato**ssico	(adj.) ‘non-toxic’
pag**ato**	PAY.p.part. ‘paid’	**ato**llo	(n.sg.) ‘atoll’
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 Experiment: micro- and macro-phonotactics

2.1 The hypothesis

Given the existence of positional regularities as a surface correlate of the morphological preference in Italian as an inflecting language (see above, §1), we are able to hypothesize that the salience of the right side of morphological complex words (i.e., the portion usually occupied by function morphemes) emerges 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 /ato/differs from similar sequences such as /tao/, /rto/, /atu/ etc.). By *macro-phonotactics*, on the other hand, we refer to positional information within the word, i.e., sub-lexical (or chunks) frequency effects (e.g., the fact that word-initial /#ato/ is different from word-medial /-ato-/ as well as from word-final /ato#/). This hypothesis was tested on a behavioural and a computational ground, within an experimental protocol aimed at correlating the speakers' responses with the computational output obtained over one and the same linguistic data set. In particular, morphologically complex pseudo-words were used to elicit similarity values (ortho-phonological similarity) from both native Italian subjects and an activation-based model trained with a phonologically encoded corpus of spoken Italian. In the following section, some details about the computational system is provided.

2.2 The Computational Model

Self-Organizing maps (SOMs) are plausible models of neural computation and learning given their sensitivity to frequency patterns in the input data and the incremental (i.e., adaptive) organization of stimuli (see Kohonon 2001).

Physically, a SOM is constituted by a topological grid of receptive fields (i.e., the neurons) which fire in presence of a stimulus or a set of stimuli (Bednar et al. 2005). By repeatedly being exposed to input stimuli, receptive fields are trained to be reactive to a particular stimulus or class of stimuli. In our simulation (Calderone et al. 2007, Herreros & Calderone 2007), receptive fields are trained on *K*-grams (in this case, 3-grams) of phonological words (corpus 'CHILDES for Italian, Calambrone section'; MacWhinney 1995). Each phoneme is binarily specified in terms of place and manner of articulation. After the training, the model outputs adjacent receptive fields for detected phonotactic regularities, i.e., for frequently attested co-occurrences of phonemes specific to the language. In other words, the receptive fields of the map develop a topographic profile of language-specific phonotactics on a distributional basis, whereby the most frequent phonotactic patterns are clearly distinguished by the neurons and quantitatively defined by an activation level which is proportional to the token-frequency of each relevant phonotactic pattern.

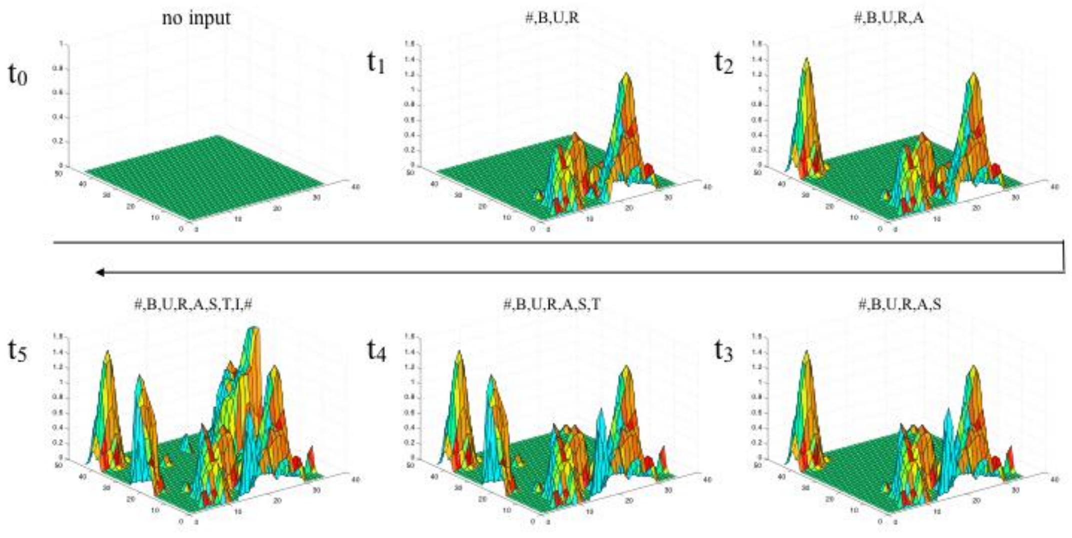
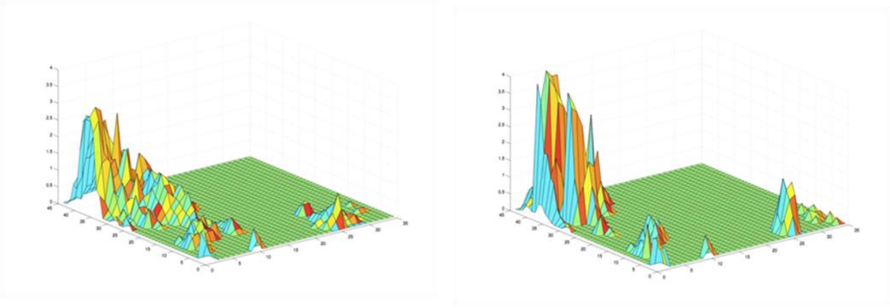


Figure 1: Activation-based lexical representation of one of the pseudo-words used in the present experiment (/burasti/), as a function of temporal progression.

Starting from this phonotactic organization of phonological stimuli, we derive a 'word representation' by means of a process of accumulation of 3-gram representations for each stimulus. In particular, the sum of the activation patterns triggered by phonological 3-grams defines a representational buffer where words are re-coded on the basis of the acquired phonotactic knowledge (Fig. 1). This means that the system performs a generalization process by summing the activation values of the3-grams, thus deriving a final vector representation of the word. The cumulative action of tri-grams' activations gives therefore a graded and distributed representation of the word in output, in which both phonological similarity (at the string level, i.e., the level of the phonological identity of segments) and token frequency effects (at the word level, i.e., the level of segments' position within the word and its frequency) are taken into account.



-ATO-
(word internal)

ATO#
(word final)

Figure 2: Activation patterns for two phonologically similar but distributionally different stimuli. Different activation magnitudes reflect differences in token-frequency among stimuli.

As a consequence, the system is able to simulate the temporary phonological storing of segmental sequences, whereby the signal is ordered and chunked to constitute higher-level units for immediate processing (Fig. 2).

Given this function of lexical representation, the overall similarity between pairs of words may be calculated in terms of the cosine distance between the two output values.

2.3 Materials and procedure

Morphologically complex pseudo-words were created by associating a non-root to an Italian inflectional or derivational affix, which was placed in either initial or final position (e.g., *burasti* vs. *stibura*). Suffixes occupied their 'legal' position when they were added to the non-root in final position (e.g., *bura**sti***), while they occupied an 'illegal' position when they were added to the non-root in initial position (e.g., ***sti**bura*); the reverse was true for prefixes (e.g., ***pre**luma* vs. *luma**pre***). Three associated items (made up of the same affix + a different non-root) were created for each pivot item (e.g., *melo**sti***, *me**sti**lo*, ***soltemi*** were associated to *bura**sti***, while ***sti**melo*, *me**sti**lo*, ***soltemi*** were associated to ***sti**bura*) (see Table 2). The three associates of each set were exactly equivalent to each other with respect to the segmental composition, but different to the extent that the affix could be placed in either the same, or a different position with respect to the affix contained in the pivot.

		Positional option	
		Initial	Final
	Pivot	stibura	burasti
Association type	Associate 1	stimelo	melosti
	Associate 2	mestilo	mestilo
	Associate 3	soltemi	soltemi

Table 2: Example of 'suffixed' pseudo-words

Both the artificial system and the pool of native Italian subjects were asked to judge the similarity of each pivot item with respect to the three associated items. The similarity ratings given by the subjects and the cosine values derived from the system were treated as independent variables and evaluated through an analysis of variance. In our hypothesis, Association type 1 should elicit higher similarity values than Association types 2 and 3 in the final positional condition more than in the initial positional condition. For example, we expect *bura**sti*** to be judged much more similar to *melo**sti*** than to *mestilo* and *soltemi*, yet the pair ***sti**bura-**sti**melo* to be judged different from ***sti**bura-mestilo* and ***sto**fera-soltemi* only to a lesser extent. In other words, we expect a significant interaction between the two independent variables of Association type and Positional option (uniformly for pseudo-words made of non-root + prefixes and of non-root + suffixes).

3 Results

3.1 Human behavior

A repeated measures ANOVA was run with Position (Initial vs. Final), Association (Type 1 vs. Type 2 vs. Type 3) and Affix (Prefix vs. Suffix) as within-subject factors. The interaction Position*Association was found to be statistically significant (Pillai's Trace, $F = 9.928$, $p < .01$),[1] while the factor Affix did not appear to affect the results of the interaction, thus indicating that pseudo-words made of prefixes and pseudo-words made of suffixes did not differ with respect to the general hypothesis. The results indicated that, in the case of the Final positional option, the Association type 1 elicited higher similarity values with respect to the two other association types, while in the case of the Initial positional option the difference between the three association conditions was not equally strong (Table 3).

		Positional option	
		Initial	Final
	Pivot	stibura	burasti
Association type	Associate 1	stimelo 5,42 (0,87)	melosti 6,29 (0,88)
	Associate 2	mestilo 4,03 (0,98)	mestilo 3,11 (0,64)
	Associate 3	soltemi 3,95 (0,77)	soltemi 2,95 (0,63)

Table 3: Results of the word similarity judgment test performed by the native Italian speakers: mean similarity ratings (s.d. in brackets) split for association type and positional option

We concluded that differences in pivot-associate relations bear different consequences when affixes are in final vs. initial position in the pivot. Given that this result was generalized to both subsets of 'prefixed' and 'suffixed' pseudo-words, the effect proved to be independent of the lexical nature of the affix. The subjects appeared therefore to be able to recover both micro- and macro-phonotactic regularities in processing complex pseudo-words.

[1] Contrasts calculated by 'difference' and by 'deviation'. Mauchlay's test for sphericity $p > .05$. Univariate within subjects test significant (Greenhouse-Geisser, $F = 7.559$, $p < .01$).

3.2 Computational simulation

A mixed design ANOVA was run with Position and Association as within-subject factors, and Affix as a between-subject factor. The interaction Position*Association was found to be non-significant (Greenhouse-Geisser correction, $F = 1.227$, $p > .05$),[2] thus indicating that the similarity differences among the three association types in initial position were equally strong than in final position (Table 4). On the other hand, the interaction Position*Association*Affix turned out to be significant (Greenhouse-Geisser correction, $F = 4.561$, $p < .05$), suggesting that 'prefixed' and 'suffixed' pseudo-words did not behave the same. Indeed, the interaction Position*Association resulted to be significant for 'suffixed' pseudo-words, non-significant for the 'prefixed' ones.

		Positional option	
		Initial	Final
	Pivot	stibura	burasti
Association type	Associate 1	stimelo 0,601	melosti 0,621
	Associate 2	mestilo 0,505	mestilo 0,522
	Associate 3	soltemi 0,485	soltemi 0,531

Table 4: Results of the word similarity rating in the output of the computational simulation: mean cosine values (s.d. in brackets) split for association type and positional option

We concluded that, in the case of the computational simulation, differences in pivot-associate relations did not bear different consequences overall when affixes were in final vs. initial position in the pivot. However, the effect was shaped by the lexical nature of the affixes composing the pseudo-words (suffix vs. prefix).

On the whole, we could say that the system is able to recover micro-phonotactic effects (thus eliciting higher cosine values for association type 1 with respect to type 2 and type 3); in addition, it appears to be able to recover macro-phonotactics effects as well, but only to a certain extent, i.e., provided that the class of the affix is specified. Indeed, the fact that the system shows higher sensitivity to segmental regularities in word final position limited to pseudo-words made of real Italian suffixes (thus being

[2] Contrasts calculated by 'difference' and by 'deviation'. Mauchlay's test for sphericity $p > .05$.

unable to abstract away the micro-phonotactic regularities from their specific segmental content) seems to suggest that, in the process of generalization, token frequency is overestimated.

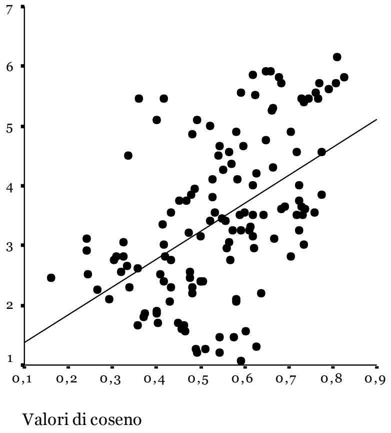


Figure 3: Global correlation between speakers' similarity ratings and computational cosine values.

3.3 Correlation human/machine

The Pearson's correlation coefficient between the observed and the simulated behaviour (r = 0.563) reported a statistically significant correlation (p < .001), thus confirming the psychological plausibility of the SOM-based simulation (Figure 3).

4 Conclusions

In conclusion, we believe that this contribution has provided some evidence in favor of the view that, in inflecting languages such as Italian, the salience of the right edge of morphological complex words (i.e., the portion usually occupied by function morphemes) may emerge as a by-product of micro-phonotactic preferences (sequential information among segments) and sub-lexical frequency effects (macro-phonotactics: positional information within the word). Positional variables, besides quantitative properties of (sub-)lexical forms, are therefore to be considered psychologically and computationally salient prerequisites for morpholexical reading.

Future work will be dealing, first of all, with the systematic introduction of phonological details in the input string used as source of the computational simulation: lexical stress has to be codified as a property of vowels/syllables, in order to achieve a closer approximation of human phonological processing of test items. Second, we will be dealing with variable windows for input data sampling (2-grams, 4-grams, 5-grams etc. besides the 3-gram sampling used in the current experiment), in order to empirically define the scope of the distributional information which is required by the system to generalize the correct information at the lexical level. Third, we would like to develop and incremental learning protocol for variable states of knowledge (i.e., variable training sets), with the explicit purpose of modeling the acquisition of morpholexical knowledge.

Bibiographical References

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

Bednar, J.A. & DePaula, J.B. & Miikkulainen, R. (2005). Self-organization of color opponent receptive fields and laterally connected orientation maps, *Neurocomputing,* 65-66: 69-76.

Boudelaa, S. & Marslen Wilson, W. (2009). Do morphological family members inhibit each other? Family size effect in Arabic, *MOPROC 2009*, Turku, Finland.

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,* Dordrecht, Kluwer, 241-264.

Bybee, J. (2007). *Frequency of use and the organization of language*, Oxford, Oxford University Press.

Calderone, B. & Herreros, I. & Pirrelli, V. (2007). Learning Inflection: The Importance of Starting Big, *Lingue & Linguaggio,* 2: 175-200.

Herreros, I. & Calderone, B. (2007). Spatial Lexical representations and Unsupervised Bootstrapping in Morphology, *Workshop on Machine Learning and Cognitive Science of Acquisition*, University College, London, June 21-22, 2007.

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

MacWhinney Brian (1995). *The CHILDES-Project: Tools for Analyzing Talk*, Hillsdale, Erlbaum (2nd edition).

McClelland & Patterson (2002). 'Words Or Rules' cannot exploit the regularity in exceptions (Reply to Pinker and Ullman), *Trends in Cognitive Science,* 6: 464-465.

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

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

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

Segalowitz, S. & Lane, K. (2000). Lexical access of function versus content words, *Brain and Language,* 75: 376-389.