The morphological impact of micro- and macro-phonotactics. Computational and behavioral analysis

Basilio Calderone*, Chiara Celata°
* MoDyCO - Université Paris Ouest Nanterre La Défense ° Scuola Normale Superiore, Pisa

1. *Introduction*

Word representation is usually modelled by postulating that phonological words can be decomposed according to either a phonotactic approach (e.g., transition probabilities) or a 'wordlikeness' approach (e.g., neighborhood density) [2]. In our approach [1,4], the phonotactic level and the word level are not conceived of as two independent domains, but rather as systems interacting in a dynamics bottom-up. In this framework, we wanted to verify whether positional variables (i.e., the occurrence of the same sound sequence in initial vs. final position within the words, all other things equal) may constitute psychological and computational significant preconditions for morphological parsing in inflectional languages such as Italian. In particular, we hypothesized that the salience of the right side of morphological complex words (i.e., the portion usually occupied by function morphemes) could emerge as a by-product of phonotactic preferences and sub-lexical frequency effects. This hypothesis was tested on a behavioural and a computational ground; a significant correlation between the two was found in response to the same set of linguistic data.

2. *Materials*

22 Italian function morphemes were selected (both inflectional and derivational, from among suffixes as well prefixes). For each affix, two sets of 4 pseudo-words (non-root morphemes + affix) were created: one pivot items, and three associated items per set. In the first set, the affix was placed in pivot initial position, while in the second it was placed in final position. Each associated item could share the pivot's affix in either final, internal or internal non-adjacent position. The three association conditions were exactly the same with respect to the segments which composed the three different pseudo-words. See Table 1 for an example.

3. *Procedure*

4.1. Humans

An off-line word similarity judgments on a 10-point scale for each pivot-associate pair was performed by 16 native Italian adult participants. Stimuli were visually presented.

4.2. Computational simulation

An unsupervised topographic map (Self-Organizing Map [1,5]) was trained with a Childes Italian phonologically transcribed corpus in order to derive phonotactic knowledge structured on a bi-dimensional grid of neurons, where similar input tokens were mapped onto neighboring output neurons. After the training, the map was able to spatially organize phonotactic sequences defined in terms of *N*-grams. Similar sequences were found in adjacent areas of the map. Each sequence was identified by a pair of coordinates in the bi-dimensional map and an activation value roughly corresponding to the frequency of occurrence in the corpus. To obtain a final representation of the word, the system performed a generalization process by summing the activation values of each *N*-gram (Fig. 1). The cumulative action of *N*-grams' activations allowed a graded and distributed representation of the word in which both phonological similarity and token frequency effects were taken into account.

An activation-based representation was derived for each pseudo-word of the corpus, and the similarity between pivot and associates was calculated in terms of the cosine distance between the two output values.

4.3. Hypothesis

Ass. Type 1 should elicit higher similarity values than Ass. Types 2 and 3, in the final positional condition more than in the initial positional condition. In other words, the interaction Association Type*Positional Option was expected to be significant.

5. *Results*

5.1.Behavioral data

The hypothesis was clearly confirmed: Ass.Type 1 elicited higher similarity values, with respect to Ass. Types 2 and 3, limited to the final positional option (F = 9,928, p < .01).

5.2.Computational data

The expected interaction proved non-significant overall (F = 1.227, p > .05), with higher similarity values for Ass. Type 1 with respect to 2 and 3 in both initial and final position. However, suffixed pseudo-words elicited higher values when the suffix was placed in final (i.e, legal) position, than when the suffix was placed in initial (i.e., illegal) position.

In conclusion, although the general hypothesis was not directly supported, the system appeared nevertheless to be able to recover, besides string-level phonotactic information, even some word-level 'paradigmatic' regularities, provided that the class of the affix was specified.

6. Correlation computational/behavioural

The Pearson's correlation coefficient (r = 0.508) reported a statistically significant correlation (p<.001), thus confirming that, in both human behavior and simulation, recurrent chunks of phonological elements are used to interpret and parse lexical items (Fig. 2).

Taken together, these data suggest that morphological salience may emerge as a by-product of distributional information at the string level and positional regularities at the word level, derived from generalizations over the inflectional nature of the language.

References

[1]Calderone/Herreros/Pirrelli (subm.) *Mapping Words: from Phonemes to Lexical Representations via Self-Organizing Map*.

[2]Bailey/Hahn 2001 *Determinants of wordlikeness: Phonotactics or lexical neighborhoods?* JMemLang 44.

[3]Albright 2002 *Islands of reliability for regular morphology: Evidence from Italian*, Lg 78: 684-709.

[4]Calderone/Celata/Herreros 2008 *Recovering morphology from local phonotactic constraints*, Laboratory Phonology 11th Conference "Phonetic Detail in the Lexicon", Wellington, NZ, 30 June - 2 July 2008.

[5]Kohonen 2001 *Self-Organizing Maps*, Springer, Berlin.

Table 1. Example of pseudo-words.

		Position	
		Final position	Initial position
	Pivot	*fera**sto***	***sto**fera*
Association Type	Association 1	*milu**sto***	***sto**milu*
	Association 2	*lu**sto**mi*	*lu**sto**mi*
	Association 3	***s**ul**t**im**o***	***s**ul**t**im**o***

Figure 1. Example of summation of N-grams for *ferasto*.

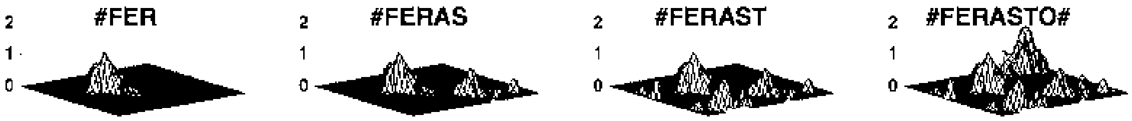
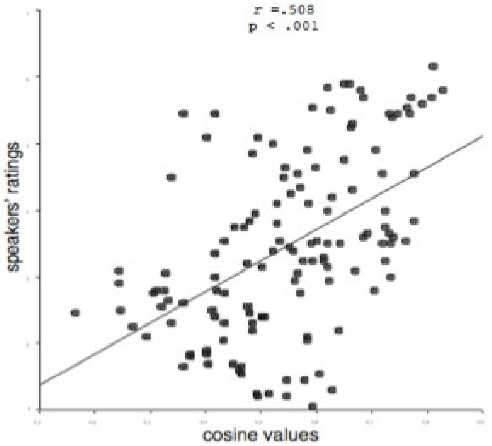


Figure 2. Global correlation between speakers' similarity ratings and computational cosine values.


r =.508
p < .001
speakers' ratings
cosine values