Phonological detail for accessing morphological structures.
Human and artificial responses in comparison

Basilio Calderone (Modyco, CNRS & Université Paris Ouest – la Défense)
Chiara Celata (Scuola Normale Superiore, Pisa)
Fabio Montermini (CLLE-ERSS, CNRS & Université de Toulouse)

1. *Introduction. The morphological impact of the quantitative properties of the lexicon*

Within the most recent models of morpholexical processing, morphemes are not recognized in isolation but rather relationally in the context of other phonologically similar material (Luce, Pisoni & Goldinger 1990, among others). In this view, morpholexical processing is to a great extent affected by statistic properties of the lexicon (or sub-parts of the lexicon), and primarily by quantitative properties of affixes, such as neighborhood density and relative frequency of morphemes (Baayen 2003; Goldsmith 2001). For example, function morphemes 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).

In Italian, a language which shows high morphological complexity while being transparent for orthography-to-phonology conversion, morpholexical reading proved an efficient and frequently activated routine for both word recognition and - more recently - word naming. It has been demonstrated that "pronunciation is obtained through activation of morphemic units both in the input and in the output lexical components", and "[in those components] the formal representations of morphemes are stored, with no necessary involvement of a semantic component as well" (Burani & Laudanna 2003: 260). 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. These differences will necessarily impact over lexical processing altogether.

2. *Micro- and macro-phonotactics: previous SOM-based simulations*

A quite basic generalization such as the one stated above can be seen as a working hypothesis to empirically test the emergent nature of morpholexical processing of complex words in natural (inflecting) languages (Bybee 2007; McClelland et al. 2002). Statistical distributions for word-final vs. non-word-final phonotactic regularities may be conceived of as a prerequisite for the emergence of paradigms and analogical rules for speakers of languages where morphology is shaped by different affixation preferences.

Previous research has shown that positional variables (i.e., the occurrence of the same sound sequence in initial vs. final position within the words, all other things equal) constitute psychological and computational significant preconditions for morphological parsing in Italian (Celata & Calderone 2010). In particular, it was verified 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 (sequential information among segments) and sub-lexical frequency effects (or *macro-phonotactics*: positional information within the word). This hypothesis was tested on a behavioural and a computational ground, within an experimental protocol aimed at correlating speakers' responses with computational outputs obtained over one and the same linguistic data set. Morphologically complex pseudo-words were used to elicit similarity values from both native Italian subjects and an unsupervised topographic map (*Self-Organising Map*, Kohonen 2002) trained with a phonologically encoded corpus of spoken Italian. 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) organisation of stimuli (see Pirrelli et al. 2004). The SOM operates on a phonotactic level by mapping similar input tokens (defined in terms of tri-gram scanning, e.g. #,T,H; T,H,E; H,E,#) onto adjacent output neurons. To obtain a final vector representation of the word, the system performs a generalization process by summing the activation values of each tri-gram. The cumulative action of tri-grams' activations gives a graded and distributed representation of the word in which both phonological similarity (at the string level) and token frequency effects (at the word level) are taken into account.

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., *fera**sto*** vs. ***sto**fera*). Three associated items (made up of the same affix + a different non-root) were created for each pivot item (e.g., *milu**sto***, *lu**sto**mi*, ***sultimo*** were associated to *fera**sto***, and ***sto**milu*, *lu**sto**mi*, ***sultimo*** to ***sto**fera*) (see Table 1). 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. 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. Results showed that the condition in which the affix position coincided between pivot and associated item crucially elicited higher similarity values, with respect to the non-coincidence conditions, for pivots with word-final affixes more than for pivots with word-initial affix (*fera**sto*** much more similar to *milu**sto*** than to *lustomi* and *sultimo*, yet ***sto**fera*-***sto**milu*** not so much different from ***sto**fera*-*lustomi* and ***sto**fera*-*sultimo*) These data were taken as evidence that morphotactic salience, preliminary to any morphological analysis, 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 inflecting nature of the language. The Pearson's correlation coefficient between the observed and simulated behaviour (r = 0.508) reported a statistically significant correlation ($p < .001$), thus confirming the psychological plausibility of the SOM-based simulation.

3. *Phonological specifications for emergent morphotactics and morphology: lexical stress*

In this study, we replicate the experimental framework to the extent that human and artificial data are elicited and compared in response to the same set of Italian pseudo-words, but some relevant changes are introduced in the domain of (1) the nature of the phonological information coded in the corpus used for SOM training, and (2) the generalization process used to derive the system's word-level representation. The two domains are strictly interconnected, inasmuch as a modification in the phonological representation of data requires specific amendment in the patterned sampling operated by the SOM-based network.

(1) In the previous experiment, words were phonologically coded following a grid of place, manner of articulation and voicing specifications (Celata & Calderone 2010 for details). In particular, vowels were specified for height and anteriority. In the present experiment, vowels are additionally specified for stress, thus distinguishing stressed vs. unstressed vowels. Given the distinctive value of lexical stress in Italian, stress specifications provide the system with a more detailed representation of the input, which is supposed to mirror the

representation of pseudo-words in native speakers' phonological competence. We consider in fact that stress pattern is part of speakers' lexical knowledge.

(2) Stress, however, is a property of (phonological) words, not just of syllables or, even less so, single phonemes. The addition of a [± stressed] feature to the set of specifications for vowels does not cope indeed with the supra-segmental dimension of lexical stress codification. For this reason, the phonotactic information previously recovered by the system by means of a tri-gram sampling of the input forms is substituted here by an algorithm of full-word memorisation achieved through sampling of larger portions of the stimulus. This associative-like lexical memorisation is then used by the system in order to produce a vector word-level representation (Figure 1).

A corpus of written Italian from the *Leipzig Corpora Collection* (Quasthoff et al. 2006) phonologically transcribed is used for the SOM training phase (word types: nearly 80,000; word tokens: nearly 5 millions). Words are phonologically coded following the specifications in (1) above. After the training, the map is able to spatially organize phonotactic sequences defined in terms of whole-word *N*-grams. Similar sequences are found in adjacent areas of the map. Each sequence is 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. Then the final representation of the word is performed through the cumulative action of *N*-grams' activations, allowing a graded and distributed representation of the word where both phonological similarity and token frequency effects are taken into account.

Similarly to the previous experiment (see above, §2), an activation-based representation is derived for each experimental pseudo-word, and the similarity between pivot and associates is calculated in terms of the cosine distance between the two output values. The values are then directly correlated to subjects' performances on the word similarity judgment task reported on in Celata & Calderone (2010) (Figure 2).

The results support the hypothesis that phonological specifications at the supra-segmental level improve the system's performance in recovering the phonological similarity of stimuli as shaped by positional, i.e., morphotactic regularities at the word level, thus providing a more accurate simulation of native speakers' performance on the same task. Differently from the previous experiment, where vowels were unspecified for stress, we find here a significant interaction between association type and affix position ($F = 5.566$, $p < .05$), thus confirming the system's ability to recover word-level 'paradigmatic' regularities, besides string-level phonotactic information. Moreover, the Pearson's correlation coefficient changes from $r = 0.508$ of the previous experiment, to $r = 0.569$, thus indicating that supra-segmental information allows the system to overlap to a larger extent human generalizations in morpholexical processing.

References

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

Burani C. & A. Laudanna (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., C. Celata & I. Herreros (2008) Recovering morphology from local phonotactic constraints, abstract edited after review for Laboratory Phonology 11th Conference (LabPhon11) «Phonetic Detail in the Lexicon»,Wellington, New Zealand 30 June - 2 July 2008.

Celata C. & B. Calderone (2010) Restrizioni fonotattiche, pattern lessicali e recupero delle regolarità morfologiche. Evidenze computazionali e comportamentali. In *Linguaggio e cervello / Semantica,* Atti del XLII Convegno della

Società di Linguistica Italiana (Pisa, Scuola Normale Superiore, 25-25 settembre 2008), Roma, Bulzoni: 47-72. Volume 2 (CD ROM).

Goldsmith J. (2001) Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics* 27(2): 153-198.

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

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

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

Pirrelli V., B. Calderone, I. Herreros & M. Virgilio (2004) Non-locality all the way through: Emergent Global Constraints in the Italian Morphological Lexicon. *Proceedings of the 7th Meeting of the ACL Special Interest Group on Computational Phonology* (SIGPHON), Barcelona: 11-19.

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

Quasthoff U., M. Richter, C. Biemann (2006) Corpus Portal for Search in Monolingual Corpora, *Proceedings of the fifth international conference on Language Resources and Evaluation*, LREC 2006, Genoa: 1799-1802.

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	***sul**timo*	***sul**timo*

Figure 1. Architecture of he SOM-based input-output mapping function.

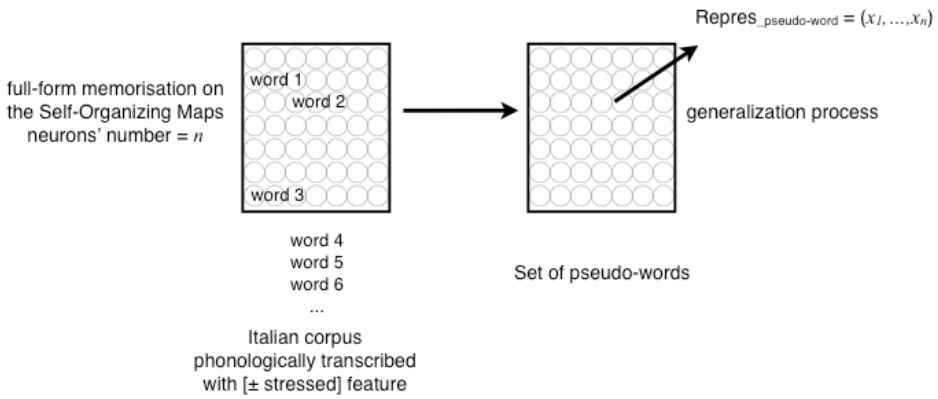

Figure 2. Global correlation between speakers' similarity ratings and computational cosine values.