Basilio Calderone
Chiara Celata
Ivan Herreros

Recovering morphology from local phonotactic constraints

LabPhon 11, Wellington, NZ, 30 June – 2 July 2008

1 Introduction

Implicit phonotactic knowledge emerges from learners' exposure to language-specific distributional information. Gradience in acceptability judgements based on phonotactic knowledge has been extensively reported on for verbal short-term memory, language processing and acquisition [1, 4]. 'Wordlikeness' studies [2] distinguish between two different types of knowledge involved in production/judgement of novel words: (a) phonotactic (string level); (b) lexical (word level). Different statistical models are used to shape the two types of knowledge (bigram/trigram transition probability for (a) and lexical neighborhood for (b) [3]).

In this work, we propose that the phonotactic level and the word level should not be conceived of as two independent, blind domains, but rather as interconnected systems interacting in a bottom-up fashion. More specifically, we test the hypothesis that the interplay of phonotactic local constraints and frequency-sensitive word patterns is powerful enough to enable over-ranked morphological patterns to be recovered from rough data.

In inflectional languages, such as Italian, function morphemes are predominantly inserted by suffixation and they tend therefore to occupy the right edge of the word. We hypothesize that, in such languages, similarity effects will emerge more salient at the right edge of a word than in other positions, all other things being equal.

2 Experiments. Materials and methods

2.1 Materials

18 Italian morphological endings were selected (both pre- and suffixes). For each morpheme, two sets of 4 pseudo-words were created: one pivot items, and three associated items per set.

In the first set (called the Final positional option), the pivot item was created by adding the relevant ending to an arbitrary root morpheme (e.g. ending: *sto*#, pivot item: *ferasto*). Each associated item could share the pivot's morphological ending in either final, internal or internal non-adjacent position (called the Association type 1, Association type 2 and Association type 3, respectively; e.g. *milusto*, *lustomi*, *sultimo*). Notice that the three association conditions were exactly the same with respect to the segments which composed the three different pseudo-words (e.g. in the case of *ferasto* and its associates, 3 matching and 4 non-matching segments). The second set of pseudo-words (called the Non-final position option) was created following the same criteria, but here the selected function morpheme was placed in word initial position (e.g. pivot item *stofera*; associated items *stomilu*, *lustomi* and *sultimo*).

In Table 1, the suffix STO is taken as an example.

		POSITION	
		Final	Initial
	Pivot	*fera**sto***	***sto**fera*
ASSOCIATION TYPE	Ass 1	*milu**sto***	***sto**milu*
	Ass 2	*lu**sto**mi*	*lu**sto**mi*
	Ass 3	***sultimo***	***sultimo***

Table 1

2.2 Methods

2.2.1 SOM-based architecture

A Self-Organizing Map (SOM) [5] is an artificial neural net which generally projects the data points of an n-dimensional input space, onto a two-dimensional output space where similar input tokens are mapped onto nearby output units.

By repeated exposure to the trigram-based training, the map develops memory traces of the most familiar phonotactic patters in an incremental way. Summation of activation patterns triggered by phonological trigrams defines a representation buffer where word forms are re-coded on the basis of acquired phonotactic knowledge.

2.2.2 Psycholinguistic task

Off-line word similarity judgments using a 10-point scale. 16 native Italian adult participants.

2.2.3 Hypothesis

The Pivot when coupled with Ass 1 should elicit higher similarity values than when coupled with Ass 2/3, and this more strongly in Final than in Initial position. In other words, the interaction between Association Type and Position was expected to be significant.

Moreover, the speakers' mean similarity rating and the SOM's cosine value, elicited on the same materials, are expected to be significantly correlated.

3 Computational evidence

The similarity values between a pivot and each of its associates was measured. The similarity between two words *a* and *b* was measured as the cosine between *a* and *b* (with values bounded between 0 and 1). The resulting cosine values were normalized w.r.t. the activation value of each form and the number of phonemes composing each morpheme.

Even if the mean similarity values were consistent with the predictions, the effect was non-significant, probably due to the paucity of data and, consequently, the high st.dev. values (see Table 2).

	Initial	Final
Pivot – Ass 1	*.211 (.063)*	*.201 (.060)*
Pivot – Ass 2	*.150 (.060)*	*.169 (.062)*
Pivot – Ass 3	*.159 (.059)*	*.163 (.058)*

Table 2

4 Psycholinguistic evidence

The hypothesis was clearly confirmed: the Association Type 1 elicited significantly ($p < .01$) higher similarity values, iff the morpheme was placed in Final Position (see Tables 3a and 3b).

	Initial	Final
Pivot – Ass 1	*5.42 (.87)*	*6.29 (.88)*
Pivot – Ass 2	*4.03 (1.13)*	*3.11 (.74)*
Pivot – Ass 3	*3.16 (.82)*	*2.95 (.73)*

Table 3a By items

	Initial	Final
Pivot – Ass 1	*5.22 (1.10)*	*6.31 (.78)*
Pivot – Ass 2	*3.55 (.75)*	*3.01 (.78)*
Pivot – Ass 3	*2.98 (.79)*	*2.35 (.87)*

Table 3a By subjects

5 Correlation

The correlation coefficient between the speakers' mean similarity rating and the SOM's cosine value was finally calculated (Fig. 1).

The *r* value proved well above chance and highly significant, at least for the majority of the morphemes.

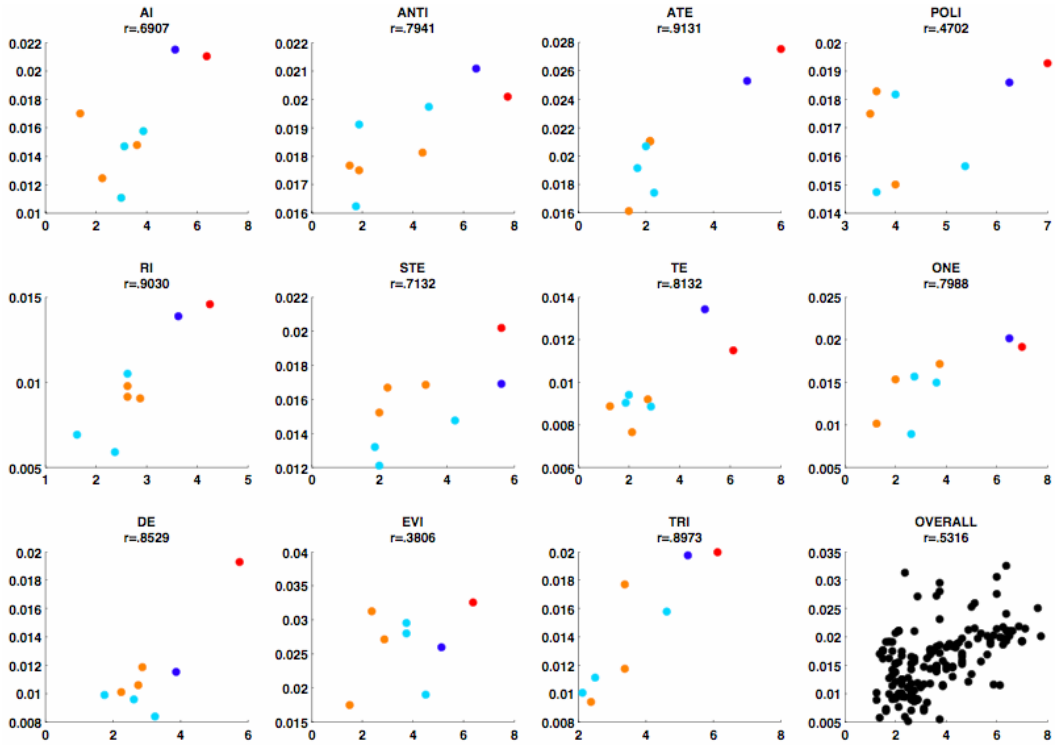

Fig. 1. Warm colors for Final position, cold colors for Initial Position; red and blue for Association Type 1, orange and azure for Association Type 2 and 3.

6 Conclusions

The interplay of phonotactic local constraints and frequency-sensitive word patterns is powerful enough to enable over-ranked morphological patterns to be recovered from rough data.

Together the SOM simulation and the behavioural evidence showed that phonotactic constraints coupled with frequency information may have significant effects at the word level, where morphological patterns emerge as the by-product of the local representations of chunks of individual elements composing the phonological chain.

7 Bibliographical references

[1] Albright, Adam. 2002. Islands of reliability for regular morphology. Evidence from Italian. *Language*, 78(4):684- 709.

[2] Bailey, Todd M. & Ulrike Hahn. 2001. Determinants of Wordlikeness: Phonotactics or Lexical Neighborhoods? *Journal of Memory and Language*, 44:568-591.

[3] Coady, Jeffry A. and Richard N. Aslin. 2004. Young children's sensitivity to probabilistic phonotactics in the developing lexicon. *Journal of Experimental Child Psychology,* 89:183-213.

[4] Coleman, John & Janet Pierrehumbert. 1997. Stochastic Phonological Grammars and Acceptability. *Third Meeting of the ACL Special Interest Group in Computational Phonology*, 49-56.

[5] Kohonen, Teuvo. 2002. *Self-Organizing Maps*, Springer, Berlin.