# Word knowledge and word usage: a Foreword

Claudia Marzi & Vito Pirrelli

*Istituto di Linguistica Computazionale "A. Zampolli", Consiglio Nazionale delle Ricerche, Pisa, Italy* <claudia.marzi@ilc.cnr.it> <vito.pirrelli@ilc.cnr.it>

This special issue, together with its companion issue to appear in *Lingue e Linguaggio*, stems from the NetWordS Final Conference *Word knowledge and word usage: representations and processes in the mental lexicon.*[*] The conference, held on the 30th and 31st of March, and the 1st of April 2015 in Pisa, concluded the 4-year NetWordS project, the European Network of Word Structure funded by the European Science Foundation within the Research Networking Programme. In line with the highly multidisciplinary profile of NetWordS agenda, the conference offered a comprehensive and inclusive forum focussing on two main lines of lexical inquiry:

(i) usage-based approaches to bootstrapping word form and structure (morpho-phonological and morpho-syntactic issues), including: acquisition of lexical categories, emergence of morphological structure, lexical memories, anticipatory prediction-based mechanisms of word recognition, word production, frequency-based models of lexical productivity, word encoding, models of lexical architecture, family-based effects in word processing, word reading and writing;

(ii) usage-based approaches to word meanings (lexical semantics and pragmatics in morphologically simple and complex words), including: distributional semantics, compound interpretation, concept composition and coercion, conceptualization of perception and action, time and space in the lexicon, metonymy and metaphor, lexico-semantic relations, perceptual grounding and embodied cognition, context-based and encyclopedic knowledge, semantic association and categorization.

The multidisciplinary focus on word knowledge and word usage promoted by the Conference led participants to openly discuss an impressive range of approaches and empirical data: priming and lexical decision in a number of contexts, distributional semantics and models of semantic composition, neural networks, machine learning and mathematical modelling of empirical evidence, as well as their neuro-biological and neuro-functional correlates.

[*] We gratefully acknowledge the European Science Foundation financial support to the NetWords final conference and this edited collection.

 *(received March 2016)*

It is widely acknowledged that looking at the same problem from different angles has an additive effect on the impact of current language research. Certainly more can be achieved, however, if, rather than simply adding more perspectives on the same subject, with individual research efforts staying within the boundaries of single knowledge domains, scholars manage to integrate them into a boundary-shifting methodological perspective. When psycholinguistic evidence from humans is successfully replicated algorithmically through a computational model implementing a few well-understood principles of time-series processing, we are in a position to empirically assess what input conditions favour memorisation and acquisition of symbolic strings by the model, and test these algorithmic predictions back on human subjects, thus going full circle. This may have a multiplicative effect on current research, providing not only mathematical modelling of present behavioural evidence, but amounting to fully explanatory mechanisms. Our current understanding of WHERE and WHEN some cognitive processes are implemented in the brain will be complemented by knowledge of WHAT information they rely on and HOW they integrate it.

Other compelling examples of the full potential of cross-disciplinary integration can be found in the present volume and in the twin issue of *Lingue e Linguaggio*. As a general point, we contend that only by putting single-domain acquisitions into the wider context of human communication, and developing an interdisciplinary framework whereby each specialist will take advantage of insights from other disciplines, we can make substantial progress in our understanding of the lexical roots of human verbal communication in real contexts. The edited selection of papers presented here provides a representative sample of the range of approaches debated at the NetWordS Pisa Conference, by way of illustration of how aspects of knowledge integration and methodological innovation can be put at the service of a better understanding of broad lexical issues.

## *The volume*

Jeroen Geertzen, James Blevins and Petar Milin in their paper *The informativeness of linguistic unit boundaries* illustrate an information-theoretic approach to assessing the amount of compressible (redundant) information conveyed by a text corpus segmented at increasing levels of granularity: from sentential, to lexical and morphological units. Starting from the standard formulation of Kolmogorov complexity in terms of Minimal Description Length, the authors find

out that the informativeness of boundaries differs across unit types and across languages (evidence is discussed for English, Estonian, Finnish and Hungarian). Unsurprisingly, text sentences are too long and too combinatorial to provide reliable redundant units for corpus compression in all languages. Sublexical constituents (morphemes) look like more promising candidate units in that respect, as confirmed by the better (higher) compression ratios obtained when morpheme boundaries are overtly marked in a text corpus for all languages. However, results on English texts show that word boundaries are even better clues for text compression, lending support to the traditional view that words are optimal-sized units for describing structural regularities in language. Finally, the richer morphology of agglutinating languages (like Hungarian and Finnish), where more pieces of morpho-syntactic information are strung together in predictable morphemic sequences, makes word boundaries result in more efficient (higher) compression rates. In a more isolating language like English, the result of word-based compression is comparatively less effective.

In *Words matter more than morphemes: an investigation of masked priming effects with complex words and non-words*, Madeleine Voga and Hélène Giraudo provide psycholinguistic support to the claim that effects of priming between morphologically related French words (like e.g. *facture* 'bill' and *facteur* 'postman') are in fact the result of associative relations holding among abstract lexical representations in the mental lexicon, rather than the outcome of morphemic splitting and morpheme-based access of morphologically complex words. Accordingly, morphologically complex words are accessed as full forms, and their morphological relations are established post-lexically, at the interface between word forms and concepts. Residual effects of priming by bound stems shown in isolation (e.g. *fact-* priming the target *facteur*) are shown to be indistinguishable from priming by orthographic controls (e.g. *bact* priming *facteur*) and thus accountable in terms of perceptual/formal effects only, rather than truly morphological relations.

A further step in the same direction is taken in *How derivational links affect lexical access: Evidence from Russian verbs and nouns*, by Natalia Slioussar and Anastasia Chuprina, showing that, in a morphologically rich language like Russian, speakers appear to be sensitive to subtle form-meaning relations between derivationally related words. According to their analysis, scope-bearing relations between members of the same derivational family appear to carve out a word family into groups of derivationally related pairs (e.g. *rodit'* 'give birth' → *rožd*e*nie* 'birth', *rodit'* 'give birth' → *porodit'* 'generate', *porodit'* 'generate' → *poroždenie* 'generation'), excluding other possible pairs which are

not linked by a direct derivational relationship (e.g. *poroždenie* is not derived from *roždenie* by prefixation). In their experiments, Russian prefixed derivatives (e.g. *porodit'*) appear to prime their bases (e.g. *rodit'*) only, not any semantically related unprefixed cognate, as shown by the failure of *poroždenie* to prime *roždenie*. Their evidence lends support to the hypothesis that, at least in some languages, only true derivational relations play a role in lexical access.

Dàniel Czégel, Zsolt Lengyel and Csaba Pléh, in *Relation between morphological and associative structure of Hungarian words*, study the correlation between two entropic scores: the morpho-syntactic entropy of Hungarian nouns and verbs (computed from a large web-based corpus) and the associative entropy of the same words, based on the frequency distribution of their lexical associates elicited from two groups of young (age 10-14) and adult (age 18-24) speakers. The hypothesis is that the distribution of the inflected forms of a word in a corpus should reflect the variety of contexts where the word occurs and, ultimately, its array of associatively related companion words. This hypothesis proves to be correct (for both age groups) for nouns only, but not for verbs, which show a reverse pattern. Authors attribute this reversal effect to the negative correlation between the number of verb associates elicited by verb targets and the associative entropy of the targets themselves (measuring the distribution of all associates, irrespectively of their part of speech), showing an age-related cognitive difficulty in organising verbal meanings through a network of hierarchical semantic relations.

*Effects of frequency and regularity in an integrative model of word storage and processing*, by Claudia Marzi, Marcello Ferro, Franco Alberto Cardillo and Vito Pirrelli, illustrates a unitary neurocomputational framework modelling neighbour family effects in serial lexical access and word recall. The model implements an integrative memory architecture, a Temporal Self-Organising Map, whereby stored lexical representations consist in routinized patterns of node activation that are repeatedly used for online processing, and processing is driven by online reactivation of successful activation patterns. In a battery of simulations run on German and Italian verb paradigms, regular forms appear to benefit from the presence of co-activating neighbours in both word recall and acquisition, where co-activation has a boosting effect and makes family members more sensitive to type frequency effects rather than token frequency effects. However, uniform distributions in large neighbour families prompt tighter competition for suffix selection, and this is in general detrimental for serial word access, where irregular and high-frequency forms are predicted more easily than regular forms.