Learning Phonotactic Preferences in Syllabification

Basilio Calderone & Pier Marco Bertinetto
(*work in progress*)

The paper investigates the use of connectionist approaches to discover phonotactic preferences in syllabification processes. By using the connectionist framework for modelling linguistic phenomena such as the syllable (e.g., [3, 4, 5, 7]), we assume that the principles governing syllabification and the internal organization of the syllable are not given as *a priori* information; rather, they emerge from the linguistic data by means of statistical regularities and categorization processes.

In our case, the syllabification task was simulated by using a feedforward two-level neural network (NN). Different simulations were implemented with Italian, Spanish and English data. Three corpora were designed, aiming at minimally representing the various syllable types typical of each language. In detail, we had the following number of words and syllable types: Italian 83 and 51, Spanish 160 and 36, English 159 and 78. The connection weights of the NN were initialized with pseudo-random values and were subsequently changed, in order to reduce the error, according to the back-propagation learning algorithm. We encoded the individual segments by using a distributed binary representation.

Every segment was defined, in this exploratory phase, on the basis of its natural class, namely as: V (vowel), G (glide), L (liquid), N (nasal), F (fricative), O (stop) and A (affricate). The input vectors for the learning protocol were created using a window encoding: every segment is coded together with its left and right phonotactic context. This methodology allows an individual segment to be computed with its complete phonological environment during the syllabification output training. After the training process (25000 sweeps, learning rate 0.4), the network exhibits correct syllabification performance for all learned word patterns.

As a second step, a generalization process was attemped. The syllabification output was obtained for 'unseen words', i.e. words not present in the training set. The NN proved to have a fairly robust knowledge, assigning syllabic boundaries to words not previously learned by extracting the relevant generalizations from the training data. New syllable types were thus discovered by the network for all corpora (Italian, Spanish and English). The system yields syllabification values ranging from 0.0 to 1.0. Needless to say, the output activation cannot reach exactly 0.0 or 1.0 as final values, but at most some approximations to the extremes of the scale (0.012 and 0.987 for example).

The above described setting appears useful for quantitative measurements relating to the phonotactic preferences of each syllabic constituent. We specially focused our analysis on the /s/ + C cluster case [1, 2, 6]. These clusters were not included in the training set of the three corpora, but the NN processed them by generalization. The Spanish and English results provide evidence for a heterosyllabic division of the /s/ + C clusters. The Italian data yielded instead a less deterministic behavior. The NN defined fuzzy values (such as 0.543 or 0.496) for Italian /s/ + C clusters, pointing towards an intermediate position between tauto- and heterosyllabic status. Experimental models, that will take into account phonotactic information and sonority features of each segment (considering the whole set of Italian consonantal phonemes), are currently under investigation.

References

[1] Bertinetto, Pier Marco. 1996. Psycholinguistic evidence for syllable geometry: Italian and beyond. In Rennison & Kähnhammer (eds.), *Phonologica 1996. Syllables!?* Holland Academic Graphics:1-28.

[2] Bertinetto, Pier Marco. 2004. On the undecidable syllabification of /sC/ clusters in Italian: Converging experimental evidence. *Italian Journal of Linguistics / Rivista di Linguistica,* 16:349-372.

[3] Goldsmith, John. 1992. Local Modeling in Phonology. In Steven Davis (ed.), *Connectionism: Theory and Practice.* Oxford, OUP:229-246.

[4] Joanisse, Marc. 1999. Exploring syllable structure in connectionist networks. In *Proceedings of the XIVth International Congress of Phonetic Sciences,* San Francisco, CA.

[5] Laks, Bernard. 1995. A connectionist account of French syllabification. *Lingua,* 95:349-372.

[6] Marotta, Giovanna. 1995. La sibilante preconsonantica in italiano: Questioni teoriche ed analisi sperimentale. In Ajello & Sani (eds.), *Scritti linguistici e filologici in onore di Tristano Bolelli.* Pisa, Pacini:393-437.

[7] Stoianov, Ivelin & John Nerbonne. 1998. Modeling the phonotactics of Natural Language Words with SRNs - part two: Exploring phonetic data representations. In *CLIN'98: Computational Linguistics in Netherlands.* Leuven, Belgium.