Processing compounds: what frequency (alone) cannot explain

Marcello Ferro1, Claudia Marzi1, Vito Pirrelli1, Christina Gagné2, Thomas Spalding3, Marco Marelli1

1Institute for Computational Linguistics CNR Pisa, 2University of Alberta, 3University of Milano-Bicocca

Abstract

Evidence on word typing [3,4] clearly indicates that morphemic structure is involved in written word production. The production of compounds differs from that of monomorphic words, and the semantic transparency of the compound constituents leads to different effects. In particular, elevation in typing latency at the morpheme boundary is larger when the first constituent is transparent than when it is opaque, but is unaffected by the transparency of the second constituent.

Furthermore, embedded pseudo-morphemes appear to influence the production of pseudo-compounds, but not in the same way that the embedded morphemes affect the production of compounds.

The evidence calls for a highly interactive processing architecture, where both compounds and constituents are processed and accessed by overlaying patterns of processing units. The level of activation of these patterns is a dynamic function of their degree of time-bound specialization for serial chunking, and their semantic contribution to the interpretation of the whole compound.

Here we show that a recurrent neural network architecture dynamically integrating these processing levels can simulate human evidence on compound typing as the outcome of competing activation patterns.

Method

A Temporal Self-organizing Map (TSOM) [2,7,9] is based on 136 compounds from G&G original data set [3]. For each compound AB (e.g. snow ball), AB, A and B are the vector-coded representations of the meanings of AB, A and B respectively, based on corpus-driven word embeddings [8]

For any AB, distance from A and B is said to estimate the amount of co-activation of A and B in the TSOM Semantic Activation Pattern. When AB is input to the map, the TSOM is then tested on the task of predicting each compound from its Semantic Activation Pattern over twenty iterations, uniformly increasing per-node activation levels by small increments at each iteration.

For letter accuracy in the task is monitored, timing the letter production latency by the number of iterations needed for each letter to be produced correctly.

For checking robustness of the effect, we repeated the same experiment with different models for semantic co-activation: dichotomous classification of compounds (opaque vs. transparent) [5], human transparency ratings [3], vector codings automatically generated by the CAOSS model [6].

Findings

- In previous work [2], TSOMs were shown to reproduce effects of structural discontinuity in the production of compounds (as opposed to monomorphic words), based on the frequency-driven specialization of re-entrant temporal connections. No effects of semantic transparency were simulated.

- Here, semantic information (word embeddings) were integrated into the architecture to simulate gradient co-activation of the meaning representations of constituents when more or less semantically transparent compounds are generated from their lexical nodes through Semantic Activation Patterns.

- Semantic transparency is shown to make compound production slower, particularly at the constituent boundary.

- Increasing semantic transparency of both constituents causes diminishing accuracy in compound production, with transparency of C1 playing a more prominent role than transparency of C2.

- Letter production latency increases at the constituent boundary, where a slowdown appears to be caused by the transparency of C1, not by the transparency of C2.

- The effect was consistently simulated using different semantic transparency scores between constituents and compounds, based on word embeddings, transparency ratings, compositional word embeddings and dichotomous classifications, thereby providing indirect support to compositional perspectives on distributional semantics [1,6].

- Linear models of letter production latency show that frequency effects are canceled out by semantic effects, reproducing evidence of human typing.

Essential references