Cognition 166 (2017) 207–224

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Original Articles

Compounding as Abstract Operation in Semantic Space: Investigating relational effects through a large-scale, data-driven computational model

Marco Marelli [a,*], Christina L. Gagné [b], Thomas L. Spalding [b]

[a] University of Milano-Bicocca, Department of Psychology, Piazza dell'Ateneo Nuovo 1, 20126 Milano, Italy
[b] University of Alberta, Department of Psychology, P217 Biological Sciences Building, Edmonton, Alberta T6G 2E9, Canada

ARTICLE INFO

Article history:
Received 26 July 2016
Revised 10 May 2017
Accepted 17 May 2017
Available online 3 June 2017

Keywords:
Compound words
Distributional semantics
Conceptual combination
Relational information
Novel compounds

ABSTRACT

In many languages, compounding is a fundamental process for the generation of novel words. When this process is productive (as, e.g., in English), native speakers can juxtapose two words to create novel compounds that can be readily understood by other speakers. The present paper proposes a large-scale, data-driven computational system for compound semantic processing based on distributional semantics, the CAOSS model (Compounding as Abstract Operation in Semantic Space). In CAOSS, word meanings are represented as vectors encoding their lexical co-occurrences in a reference corpus. Given two constituent words, their composed representation (the compound) is computed by using matrices representing the abstract properties of constituent roles (modifier vs. head). The matrices are also induced through examples of language usage. The model is then validated against behavioral results concerning the processing of novel compounds, and in particular relational effects on response latencies. The effects of relational priming and relational dominance are considered. CAOSS predictions are shown to pattern with previous results, in terms of both the impact of relational information and the dissociations related to the different constituent roles. The simulations indicate that relational information is implicitly reflected in language usage, suggesting that human speakers can learn these aspects from language experience and automatically apply them to the processing of new word combinations. The present model is flexible enough to emulate this procedure, suggesting that relational effects might emerge as a by-product of nuanced operations across distributional patterns.

1. Introduction

Compounds (e.g., *boathouse*) are morphologically complex words consisting of the concatenation of two or more independent lexical entries (the constituents; in our example, *boat* and *house*) and are the result of one of the most productive and widespread procedures for word combination. As such, they are also one of the main mechanisms for lexical enrichment (Downing, 1977): through simple concatenation of existing words, compounding permits the generation of new lexical entries that are easily understood by language speakers, even at first encounter. The present paper uses computational techniques as a way of providing a better understanding of this remarkable human ability. We will present a new large-scale model for the semantic processing of novel compounds, trained on examples of natural language usage through methods from distributional semantics (Landauer & Dumais, 1997), and test it against results from experimental psychology. This introductory section will frame our proposal in the wider context of the previous psycholinguistic and computational literature. First, we will briefly present experimental results concerning the processing of novel compounds, and draw a theoretical link between two different research traditions, namely morphological processing and conceptual combination. Second, we will discuss previous attempts at modelling compound-word meanings through computational methods, and highlight the aspects we consider of particular interest for our purposes.

1.1. From combinations of words to combinations of meanings

Given compounding's central linguistic role, it is not surprising that compounds have received much attention in psycholinguistics, and in particular in the morphological processing community.

* Corresponding author.
E-mail addresses: marco.marelli@unimib.it (M. Marelli), cgagne@ualberta.ca (C.L. Gagné), spalding@ualberta.ca (T.L. Spalding).

http://dx.doi.org/10.1016/j.cognition.2017.05.026
0010-0277/

Since the seminal study by Taft and Forster (1976), the research has been driven by questions concerning the role of constituents in word processing: is a compound word represented and processed as an independent unit, or do constituent representations mediate compound access? Results in favor of a constituent role in compound processing accumulated during the four decades after Taft and Forster's (1976) study, spanning several languages and experimental paradigms. For example, the influence of constituents is crosslinguistically well-established in the constituent-priming literature: responses to a compound word are facilitated (i.e., shorter with respect to a control condition) when the compound is preceded by one of its constituents. Such results were reported for a number of different languages (Dutch: Zwitserlood, 1994; English: Libben, Gibson, Yoon, & Sandra, 2003; French and Bulgarian: Jarema, Busson, Nikolova, Tsapkini, & Libben, 1999; Greek and Polish: Kehayia et al., 1999; Italian: Marelli, Crepaldi, & Luzzatti, 2009), and the effect holds across different manipulations (e.g., considering compounds with a common constituent as prime-target pairs: Duñabeitia, Laka, Perea, & Carreiras, 2009; varying stimulus-onset-asynchrony: Fiorentino & Fund-Reznicek, 2009). Further supporting evidence is found in paradigms requiring less heavy manipulation, showing that constituent-related properties have an impact on compound processing that is largely task independent. For example, the frequency of the constituents *as independent words* modulates response latencies in simple lexical decision (e.g., Duñabeitia, Perea, & Carreiras, 2007), latencies in word naming (e.g., Juhasz, Starr, Inhoff, & Placke, 2003), and fixation times in reading (e.g., Hyönä & Pollatsek, 1998). Moreover, family size (i.e., the number of complex words including a given constituent) has also been shown to affect processing in both behavioral (e.g., Baayen, 2010; De Jong, Feldman, Schreuder, Pastizzo, & Baayen, 2002) and eye-tracking measures (e.g., Kuperman, Schreuder, Bertram, & Baayen, 2009; Kuperman, Bertram, & Baayen, 2008). In conclusion, although the results may not be as clear in word production (compare spoken production, Janssen, Pajtas, & Caramazza, 2014, with written production, Gagné & Spalding, 2016), and modelling details are still debated (Pham & Baayen, 2015), the research indicates that constituents play a role in the processing of compound words.

Several proposals have been advanced to account for the role of constituents in compound processing. According to Libben (1998) representations of constituent words are entangled with the compounds they belong to, at both a lexical and a semantic level. More recently, Libben (2014) framed the relationship between a compound and its constituents as part of a highly interconnected system. Crucially, this system includes separate representations for words as independent units and for words as positionally-bound compound constituents. The latter are generated through a procedure of lexical and semantic drift, driven by language usage, and which facilitates compound access by structuring the complex and rich network of lexical representations. Conversely, Kuperman (2013) suggests that constituent-related effects might be the expression of a discriminative process between compound and constituent meaning representations. Orthographic cues in the visual input will always activate both the compound and its constituents, but the extent to which these representations compete with each other will depend on previous language-learning experiences (see also Pham & Baayen, 2015).

A third explanation of the constituent-related effects finds its foundations in an active compositional procedure aimed at generating the compound on the basis of the processed constituents (Gagné & Spalding, 2004, 2009). This proposal rests on the observation that compounds are not only concatenations of words, but also *combinations of meanings*: a *doghouse* is a *house* for *dogs*. This combination is based on a (most often) well-defined hierarchical structure (Di Sciullo & Williams, 1987; Williams, 1981), where the rightmost constituent takes the head role, specifying the category and most prominent lexical and semantic features of the compound (a *doghouse* is a *house*), whereas the leftmost constituent is the modifier, providing finer-grained specification to the combined meaning (a *doghouse* is something meant for *dogs*). Results from psycholinguistic experiments indicate that a compositional procedure building on these constituent roles is a routine operation during compound processing. This is particularly evident from studies investigating compound semantic transparency. Indeed, it is not always possible to deduce a compound meaning on the basis of its constituents: different degrees of transparency can be observed, ranging from very transparent cases where the compound is easily understandable on the basis of its constituents (e.g., *doghouse*), to very opaque cases where the compound meaning is rather arbitrary (e.g., *hogwash*). The effect of semantic transparency is most evident when operationalized in compositional terms either through experimental settings (El-Bialy, Gagné, & Spalding, 2013; Frisson, Niswander-Klement, & Pollatsek, 2008; Ji, Gagné, & Spalding, 2011), rating-task instructions (Marelli & Luzzatti, 2012), or modelling details (Marelli, Dinu, Zamparelli, & Baroni, 2015). Results from these studies indicate that an opaque compound is more difficult to process (e.g., it evokes longer latencies), which is arguably a consequence of a conflict between the compound meaning computed through the combinatorial procedure, and its idiosyncratic meaning stored in memory. In other words, for an opaque compound a combinatorial procedure will necessarily produce a "wrong" combined concept, that will be at odds with the one learnt through language experience and dependent on lexical knowledge; on the other hand, the combinatorial procedure will assign to transparent compounds meanings that are close to their lexicalized ones, making the transparent word easier to process.

The pervasiveness and automaticity of a combinatorial procedure might sound surprising, especially considering how it can even hinder processing when it creates erroneous outcomes. However, the routine application of such a procedure is actually very plausible when considering that compounding is, first and foremost, a mechanism for creating and understanding novel forms. Native speakers have clear intuitions concerning the meaning of compounds encountered for the very first time, and an automatic compositional procedure explains how this is possible. In turn, the importance of such a procedure for language understanding accounts for its pervasiveness; because there is no way to know in advance whether a word combination is familiar or not, the compositional procedure will end up being applied also to known compounds (e.g., El-Bialy et al., 2013). Granted, the processing of lexicalized compounds and that of novel combinations will always crucially differ in terms of its final result: whereas the former must eventually *access* a previously stored representation (the familiar meaning of the compound word), the latter aims at *generating* a novel meaning for the unfamiliar combination, building on the familiar representations of the constituent words. In the latter case, the constituents can only contribute to the compound meaning; in the former case, in addition to the composition process, the constituents can also compete with and/or facilitate the representation of the familiar compound. However, given the centrality in both processes of the compositional mechanism, investigating it is a crucial step towards the understanding of compound-word processing in general (e.g., Gagné & Spalding, 2014a).

Compositional procedures have been thoroughly explored by the literature on conceptual combination, which has shown how the ability to combine concepts in novel ways underlies creativity, allowing one to explore new thoughts and imagine new possibilities - novel compound words are the linguistic form by which these new thoughts can be expressed (Gagné & Spalding, 2007). According to the conceptual combination literature, understanding