Measuring distributional semantic effects in syntactic variation

Kristina Gulordava
Computational Learning and Computational Linguistics group
Linguistics Department, University of Geneva

Introduction and previous work  A number of papers in the quantitative syntax literature have proposed that semantic dependencies between words affect their relative order in a sentence [13, 15]. This idea is an extension of the very general principle stating that (syntactically) related words tend to occur next to each other [4]. This principle is known under various names in syntax and in the language processing literature, such as early immediate constituents, locality, and dependency length minimisation [12, 8, 18, 7]. Semantic and syntactic relations between words are highly correlated notions and it is difficult to tease apart the syntactic from the purely semantic effects on word adjacency. Lohse et al. (2004) [15] unambiguously identifies a semantic pressure favouring adjacent orders in verb-particle construction in English. A verb-particle transitive phrase can have two alternative word orders: verb-particle split order (1) and verb-particle adjacent order (2).

(1) He threw a bag of trash out.
(2) He threw out a bag of trash.

It is well known that the size of the object NP affects speaker preferences (and relative corpus frequencies) between these two orders [12, 9]: the longer the NP, the less probable the split order (1). Importantly, Lohse et al. (2004) demonstrate that there is an additional effect of the verb-particle semantic dependency. More precisely, based on off-line questionnaire responses, they classify the semantic relation between verbs and particles using two dependencies: whether the verb meaning is dependent on the particle (in other words the meaning of verb+particle phrase does not entail the meaning of the verb) and whether the particle meaning is dependent on the verb (the meaning of the verb+particle phrase does not entail the meaning “be/become”+particle). In their corpus-based experiment, Lohse et al. (2004) found that semantically-dependent particles occur in a split order, such as (1), much less frequently than independent particles. Dependent verbs, on the other hand, showed the same relative preferences for the two word orders as independent verbs.

General goals  In our work, we follow Lohse et al. (2004) in the pursuit of semantic dependency effects in syntactic variation and extend their approach in two ways. First, we investigate a different type of construction which exhibits word order alternations between adjacent and non-adjacent orders: the placement of adjectives in complex noun phrases in Romance languages. Second, Lohse et al. (2004) have addressed the semantic dependency (or entailment relation) in categorical terms since they manually annotated the verb-particle pairs with entailment judgments. We assume, instead, that this relation is not categorical and that the semantic dependency “strength” can be accessed using quantitative measures proposed previously in distributional semantics [1, 16, 5]. The overall goal of this novel line of work is therefore to establish an empirical and theoretical connection between the distributional semantics of words and their syntactic variation.
Experimental details  Romance languages show a substantial amount of adjective-noun variation. For some adjectives, we can consider the two possible orders — prenominal and postnominal — to have the same meaning as in ‘un compito$_N$ importante$_A$’ and ‘un importante$_A$ compito$_N$’ (Italian, ‘important homework’). Examples (3-5) illustrate possible adjective placements in complex noun phrases which can include, for example, a prepositional phrase as a complement of the noun.

(3) Oggi devo finire [un importante$_A$ compito$_N$ di matematica] e quindi non posso uscire.
(4) Oggi devo finire [un compito$_N$ importante$_A$ di matematica] e quindi non posso uscire.
(5) Oggi devo finire [un compito$_N$ di matematica importante$_A$] e quindi non posso uscire.

‘Today I have to finish an important math homework so I can’t go out.’

In this work, we will measure how the semantic relation between the noun and its complement (‘compito di matematica’) affects the choice between the three possible word orders. A hypothesis, analogous to the verb-particle split construction, is that when there is a strong semantic dependency between the PP-dependent and the noun, the N-PP adjacent orders (3) and (5) will be preferred compared to the order (4), where the adjective intervenes between the two elements. An intuitive motivation for this hypothesis comes also from the fact that compounds such as ‘nuovo (*?fresco) di Pasqua’ are not separable, while noun-adjunct phrases such as ‘nuovo (fresco) della fattoria’ can be non-adjacent. Our investigation aims to reveal a gradual dimension between these two cases.

However, the quantitative study of distributional semantic effects on syntactic variation is a novel line of research and we first will focus on defining relevant ‘semantic dependencies’ in N-PP phrases. Candidate lexical relations well-studied in distributional semantics include non-compositionality (symmetric relation) and lexical entailment (asymmetric relation, used in Lohse et al. (2004) study). To quantify the “strength” of the semantic dependencies we will choose among the corpus-based distributional measures developed in the previous work [1, 6, 19, 14, 17]. We will then test these semantic measures against the counts of orders (3-5) observed in the complex noun phrases extracted from dependency-annotated corpora. Note that the semantic measures are computed from corpus counts which we will collect to be disjoint from the adjective placement counts. The observed variable $sem(N, PP)$ will therefore be computed independently from the dependent variable $order$.

Previous work [11, 10] has extracted the cases of adjective-noun variation from the gold-annotated dependency corpora. In this work, we will extend these data using much larger automatically annotated treebanks. Starting with very large amount of annotated data is necessary since complex noun phrases including adjectives and a noun complement are not frequent. At the moment, we have used the Italian parsed Wikipedia [2] and extracted around 2’700’000 cases of noun phrases with at least one adjective and a noun dependent (for example, out of these, around 300’000 cases include a PP-complement with the preposition $di$). The subset of instances with adjective types appearing both prenominally and postnominally includes 560’000 cases (out of these, around 20’000 instances are noun phrases with a $di$-complement). We later plan to add the data from other Romance languages to our analysis. As in the previous work [10], we will use Generalised Mixed Effects modelling [3] to statistically test whether the proposed semantic dependency measures affect the adjective placement across the corpus, controlling for adjective lexical variation through the use of random factors.
References


