The different meanings of 'a': Capturing *qualia* relations of Italian complex nominals with distributional semantics

Sandro Pezzelle
University of Trento
sandro.pezzelle@unitn.it

Elisabetta Ježek
University of Pavia
jezek@unipv.it

Maria Silvia Micheli
University of Pavia
m.micheli1@studenti.unibg.it

This paper examines the semantic role of the preposition 'a' in *NaN* Italian complex nominals using a distributional semantic approach. Starting from the assumption that 'a' may introduce one the following *qualia* relations (Pustejovsky, 1995) - Formal (F, introducing taxonomic information, such as shape in *cacciavite a stella* 'star screwdriver'), Constitutive (C, introducing information on parts, as in *codice a barre* 'barcode'), Telic (T, introducing information on purpose and function, as in *barca a vela* 'sailboat') - we verified whether the difference in the semantic contribution of 'a' in T ('a-telic'), F ('a-formal') or C ('a-constitutive') *NaN*s is confirmed by a semantic analysis performed using vector models. We generated meaning representations for each preposition 'a' using a distributional semantic approach. First, we extracted all *NaN*s with frequency > 5 from the 1.7B tokens itWaC corpus (Baroni et al., 2009). Then, two of the authors annotated them with F, C, or T according to the scheme in Bouillon et al. (2012). In total, annotators agreed on 66 *NaN*s (19 C, 21 F, and 26 T). Finally, we generated meaning representations for both *NaN*s and single *N*s by training a `word2vec` model by Mikolov et al. (2013) on the whole corpus. Meaning representations for each preposition 'a' were obtained by subtracting the vector resulting from the sum of the nouns (e.g. *barca*+*vela*) from the *NaN* vector *barca_a_vela*). The resulting vectors were then used for running a cluster analysis. With 3 clusters, 'a-telic' clustered together (78%), with 'a-formal' forming a relatively defined cluster (52%) and 'a-constitutive' being almost equally distributed among the clusters. With 2 clusters, the distinction turns out to be much clearer, with 'a-telic' items (76% in cluster 1) clearly distinguished from the 'a-non-telic' (83% in cluster 2). Interestingly, all the 'a-non-telic' clustered with 'a-telic' are constitutive.

References: • Baroni M. et al. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. • Bouillon P. et al. (2012). Annotating qualia relations in Italian and French complex nominals. • Mikolov T. et al. (2013). Efficient estimation of word representations in vector space. • Pustejovsky J. (1995) The Generative Lexicon. Cambridge, MIT Press.