Semantic Vector Space Model and the Usage Patterns of Indonesian
Denominal Verbs with meN-, meN- -kan, and meN- -i Affixes

Gede Primahadi-Wijaya-Rajeg a, Karlina Denistia b, and Simon Musgrave a
Monash University, Australia a & Eberhard Karls University of Tübingen, Germany b

Verbal derivation takes a central position in the description and analysis of the Indonesian language. This paper proposes an approach using a Semantic Vector Space (SVS) model (e.g., Erk, 2012) to explore the semantic (dis)similarity between a set of denominal verbs derived with meN-, meN- -kan, and meN- -i affixes. Our study investigates whether verbs, such as melangkah ‘to step’, melangkahkan ‘to step up’, and melangkahi ‘to step over’, of the same noun-root (i.e., langkah ‘(a) step’) exhibit distinct or similar co-occurrence usage patterns in a corpus. We report here on which word(s) of a given derivational set pattern differently and what their co-occurrence patterns (e.g., collocations) reveal regarding this difference.

We built an SVS model with the wordVectors R package (Schmidt & Li, 2017) on the basis of the whole files in the Indonesian Leipzig Corpora (Goldhahn, Eckart, & Quasthoff, 2012). SVS represents words as points in multidimensional space, reflecting the co-occurrence patterns of each word in the corpus. Words with similar vector-space values along these dimensions are closer in semantic space (Erk, 2012). Based on this model, and by means of MorphInd (Larasati, Kuboh, & Zeman, 2011), we examined the denominal verbs, having a minimum token frequency of ten and occurring with all three affixes. We first measured the Cosine Distances between the verbs; the smaller the distance between the verbs, the closer their semantics. Then, we performed Hierarchical Agglomerative Clustering (HAC) analysis on the distances. We expect that verbs with similar co-occurrence patterns would cluster together. Figure 1 shows that for most verbs the base form (i.e., the meN- form) and the two derived forms (i.e., meN- -i and meN- -kan) do cluster together. There are no cases where the base form does not cluster with one of the derivatives, but there are several bases for which one derived-form separates from the other two forms. The separated form can be either the -i derivative (e.g., membuahi ‘to fertilise’) or the -kan derivative (e.g., mencontohkan ‘to exemplify’).

The most semantically coherent group consists of foot-related motion verbs based on the nouns langkah ‘step’, tapak ‘sole of the foot’, and jejak ‘footprint’ (see the cluster at the very bottom of the dendrogram in Figure 1). Even in this case, the meN- -i form of langkah (i.e., melangkahi ‘to step over’) is in different cluster. Informal inspection on the collocational data of melangkahi ‘to step over’ reveals that it is most frequently used in a metaphorical sense related to ‘obligation’, given the more abstract nature of its direct objects (e.g., aturan ‘regulation’, kewenangan ‘authority’, and tugas ‘task’). These co-occurrence patterns potentially differentiate melangkahi from melangkahkan ‘to step up’, despite both being transitive verbs; melangkahkan frequently collocates with kaki ‘foot’ as its direct object. The intransitive melangkah ‘to step; tread’ is in a cluster with melangkahkan presumably because they mostly share animate subject collocate such as a third-person singular pronoun; these co-occurrence patterns suggest the more ‘concrete motion’ sense of melangkah ‘to step’ and melangkahkan ‘to step up’.

Summing up, we provide a usage-based overview on denominal verbs with meN-, meN- -kan, and meN- -i. Different clustering patterns between these affixes indicate differences in the resulting co-occurrence patterns of the verbs the affixes derive. We will expand the potentials of SVS to find semantic clusters for (i) verbs of the same noun root but of different voice prefix (e.g. comparing the active meN- with the passive di-, provided the verbs do occur in the di-form), and for (ii) verbs with roots of other word classes (e.g., adjective and verb).
Hierarchical Agglomerative Clustering (HAC) analysis on words with 'Noun' root.  
(Distance measure = 'Cosine'; Clustering method = 'ward.D2')

Figure 1 HAC dendrogram on the Indonesian denominal verbs with meN-, meN- -kan, meN- -i affixes.

References