Table 3.1 Example of a real cluster (421_n) in the Diana-Araknion corpus in Spanish

Cluster: 421_n			
Lemmas (c_{421_le})	barba_n, bigote_n, cabellera_n, cabello_n, ceja_n, crin_n, melena_n, mostacho_n, patilla_n, pelaje_n, pelo_n, perilla_n, vello_n		
Contexts (c_{421_ctx})	[< : dobj : erizar_v],11	[< : oth : erizar_v],11	[< : oth : rizar_v],10
	[< : subj : erizar_v],10	[> : mod : espeso_a],9	[> : oth : espeso_a],9
	[> : mod : negro_a],7	[< : oth : negro_a],5	[> : mod : gris_a],8
	[< : dobj : rizar_v],8	[> : oth : gris_a],7	[< : oth : pelo_n],6
	[> : mod : rubio_a],7	[> : mod : barba_n],7	[< : oth : atusar_v],7
	[> : mod : largo_a],4	[> : oth : rubio_a],6	[< : mod : pelo_n],2
	[> : mod : rojizo_a],4	[> : oth : rojizo_a],6	[> : oth : largo_a],3
	[< : oth : bigote_n],3	[> : mod : blanco_a],3	[> : mod : cano_a],5
	[> : mod : hirsuto_a],5	[> : oth : hirsuto_a],2	[> : oth : largo_a],3
	[> : oth : negro_a],2	[> : mod : rojizo_a],2	

3.3.4.1 Results of the Clustering Process

Following our configuration, we obtained a total of 1,500 clusters in the clustering process (k=1500). It is worth noting that the clusters are highly morpho-syntactically and semantically cohesive.

The clusters contain lemmas belonging mostly to the same POS. It is worth mentioning that more than half of the clusters are nouns (54.20%), followed by verbs (25.80%) and adjectives (16.67%). Clusters of adverbs make up only 3.33% of the total.

Clusters contain relevant implicit information, in the sense that their lemmas belong to well-defined semantic categories, often at a very fine-grained level. For instance, we obtained clusters of adjectives with a *Positive Polarity* (5) and with a *Negative Polarity* (6)[26]. These results encourage us to tag all the clusters with one or more semantic labels. That will enrich the obtained patterns.

5. {c_{111}, *Positive_Polarity* adjectives: admirable_a, asombroso_a, genial_a...}[27]

6. {c_{38}, *Negative_Polarity* adjectives: atroz_a, aterrador_a, espantoso_a...}[28]

[27] 'admirable, amazing, great'
[28] 'atrocious, scary, frightening'

3.3.5 Generalization: Linking and Filtering Clusters

The process of generalization by linking clusters (see Step 4 in Figure 3.1) is based on the set of clusters and contexts obtained using CLUTO. The processes of linking clusters and pattern generation detailed in Section 3.3.6 are the core steps of the DISCOver methodology. The process of linking clusters uses the set of the twenty-five highest scored contexts in each cluster. According to our pattern-construction hypothesis (see Section 3.3.1), the goal of the linking of clusters is to establish the relationships between clusters using their contexts, as defined in (3.3), obtaining as a result a matrix of all possible contextual relations between clusters (see Section 3.3.5.1). Next, we apply a filtering process in order to select strongly related links taking into account different criteria (see Section 3.3.5.2).

3.3.5.1 Linking Clusters and Building the Matrix of Related Clusters

Basically, the aim of the cluster linking process is to establish the relationships between clusters and to store them in a matrix, $R\_clusters$, with k rows and k columns. The k-value corresponds to the number of clusters obtained in the clustering step.

For building the matrix, for each origin cluster (x) each dep_dir and dep_lab of the $context\_cluster$ (defined in Equation 3.3) are converted into a $contextual\_relation$ (see Equation 3.4), while the $context\_lemma$ of the $context\_cluster$ is used to locate the cluster (y) in which it occurs. We obtain as a result a matrix, $R\_clusters$, in which clusters are related according to a set of contextual relations stored in a $relation\_set$. The sum of the scores of the $context\_clusters$ in 3.3 are added together in a matrix, R_scores. The R_scores matrix is later used in the process for determining filtering thresholds.

$$contextual\_relation =< dep\_dir, dep\_lab > \tag{3.4}$$

For the contextual relation, defined in 3.4, dep_dir and dep_lab are the dependency direction and the dependency label defined in a context of cluster i related to cluster j. Note that the $relation\_set$ of a cluster in itself is empty as $R\_clusters[i][i] = \emptyset$ and $R\_clusters[i][j] \neq R\_clusters[j][i]$.

Following the example of cluster 421_n, described in Table 3.1, the result of the cluster linking process for this particular cluster ($i = 421\_n$) is shown in Table 3.2[29]. The first column in this table shows the related clusters, j, the second column shows the relation_type that relates cluster 421_n to the related clusters j (i.e. STRONG, SEMI or WEAK, See 3.3.5.2), and finally the last column describes the lemmas in the related clusters.

[29]For the sake of simplicity, the contexts are not included in the Table 2 and we only show a relation of each type.

Table 3.2 Some examples of cluster linking process in cluster i=421_n (described in Table 3.1).

Related clusters(j)	**Relation_ type**	**Lemmas** ($c_{j.le}$, where c_j refers to the related cluster, j)
1223_a	STRONG	azabache_a, bermejo_a, **cano_a**, canoso_a, **hirsuto_a**, lacio_a, lustroso_a, ondulante_a, sedoso_a...
932_v	SEMI	afeitar_v, **atusar_v**, cepillar_v, empolvar_v, enguantar_v, peinar_v, rasurar_v...
405_n	WEAK	contario_n, final_n, largo_n, menudo_n...

3.3.5.2 Filtering Related Clusters

In the $R\_clusters$ matrix, not all contextual relationships between clusters are accepted since they have a low R_scores. For this reason, we established two criteria to automatically determine which relationships will be maintained and which ones are filtered out in the pattern generation process. For each criterion only those relations higher than a predetermined score value will be considered. The criteria are the following:

- **Criterion 1**: For each pair of clusters i and j, we take into account those relations that in each of their directions (i.e., $R\_scores[i][j]$ or $R\_scores[j][i]$) have a score above a minimum predetermined value, that is, $threshold_1$. This $threshold_1$ is automatically determined by finding a score value that allows for the grouping of 30% of the clusters. The relations that fulfill criterion 1 are called STRONG relations.

- **Criterion 2**: For each pair of clusters i and j, we take into account those relations in which the sum of scores in both directions (i.e., $R\_scores[i][j]+$ $R\_scores[j][i]$) is higher than a predetermined value, that is, $threshold_2$, which is determined by finding a value that allows for the grouping of 50% of the clusters. The relations that fulfill criterion 2 are called SEMI relations.

Considering the example of cluster 421_n, the result of the filtering process is that, out of the three clusters linked to cluster 421_n in our example[26] (1223_a, 932_v, and 405_n), we will only select those with STRONG and SEMI relations, that is, 1223_a, and 932_v. Those labelled as WEAK (e.g., 405_n shown in Table 3.2) are filtered out because they do not reach the established thresholds.