### 3.3.6 Pattern Generation

Once the process for automatically linking and filtering clusters was carried out, we proceeded to generate the lexico-syntactic patterns to be considered as candidates for constructions (see Step 5 in Figure 3.1). Each generated pattern is defined as follows:

$$pattern =< lemma_i, dep\_dir, dep\_lab, lemma_j > \tag{3.5}$$

where $lemma_i$ and $lemma_j$ are the lemmas contained in the related clusters ($i$ and $j$), $dep\_dir$ and $dep\_lab$ are the dependency direction and the dependency label between the related clusters. So, there is a pattern for each $lemma_i$ and $lemma_j$ pair.

As we mentioned in Section 3.3.4, all possible configurations using between 10,000 and 15,000 lemmas gave acceptable related clusters. In order to increase the number of patterns generated we carried out the same process with a configuration using 10,000 lemmas. We combined the patterns obtained using the 10,000 and 15,000 lemmas together and removed those that were shared by both configurations. In Tables 3.3, 3.4 and 3.5, we show the number of resulting clusters and patterns, after removing the overlapping patterns, for the two configurations.

**Table 3.3** Distribution of the number of related and unrelated clusters and their percentage

| | **10,000 lemmas** | **15,000 lemmas** |
|---|---|---|
| **Relation** | **Clusters (%)** | **Clusters (%)** |
| STRONG | 441 (31.50%) | 461 (30.73%) |
| SEMI | 339 (24.21%) | 396 (26.40%) |
| **Total** | **780** (55.71%) | **857** (57.13%) |
| | | |
| WEAK | 589 (42.07%) | 636 (42.40%) |
| Unrelated | 31 (2.21%) | 7 (0.47%) |

As shown in Table 3.3 (second and third columns), more than 55% of the linked clusters maintain STRONG and SEMI relationships, whereas only the 2.68% of the clusters remain unrelated. Table 3.4 (second and third columns) shows the distribution of linked clusters by POS in both configurations.

The total number of lexico-syntactic patterns obtained from the two configurations of clusters (780 and 857 STRONG and SEMI related clusters) is 237,444. For the purpose of pattern generation, STRONG and SEMI clusters have been treated equally. From these patterns, we removed 16,712 patterns, those that were present in both sets of generated patterns, given as a result the total number of 220,732 patterns (See Table 3.5).

**Table 3.4** Distribution of the number of related clusters and their percentage by POS

| | **10,000 lemmas** | **15,000 lemmas** |
|---|---|---|
| **POS** | **Clusters (%)** | **Clusters (%)** |
| N | 415 (53.21%) | 464 (54.14%) |
| V | 197 (25.26%) | 182 (12.24%) |
| A | 142 (18.21%) | 173 (20.19%) |
| R | 26 (3.30%) | 38 (4.43%) |
| **Total** | **780** (100%) | **857** (100%) |

**Table 3.5** Distribution of the generated patterns

| **Lemmas** | **Attested-Patterns** | **Unattested-Patterns** | **Total** |
|---|---|---|---|
| 10,000 | 23,980 | 48,147 | 72,127 |
| 15,000 | 37,840 | 127,477 | 165,317 |
| 10,000 + 15,000 | 61,820 | 175,624 | 237,444 |
| **Overlapping** | **8,531** | **8,181** | **16,712** |
| **Sum (no overlap)** | **53,289** | **167,443** | **220,732** |

The DISCOver methodology allows for the generation of patterns that actually occur in the corpus (Attested-Patterns), but also of lexico-syntactic patterns that are not present in the corpus but which are highly plausible in Spanish (Unattested-Patterns), since the components of the clusters are closely semantically related. As a result, we are able to enlarge the descriptive power of the source corpus. Among the patterns we generated, 61,820 were Attested-Patterns, that is, patterns that are present in the source corpus, and 175,624 were Unattested-Patterns, that is, new patterns (see Table 3.5).

Retaking the example of cluster 421_n and its related clusters we obtain patterns such as those shown in (7)[30]:

7. <bigote$_{c_421}$ <:dobj: cepillar$_{c_932_v}$>
   <melena$_{c_421}$ <:dobj: alisar$_{c_1267_v}$>
   <pelaje$_{c_421}$ >:mod: sedoso$_{c_1223_a}$>
   <perilla$_{c_421}$ >:mod: gris$_{c_149_a}$>

All of these patterns are Unattested-Patterns, that is, they do not occur in the Diana-Araknion corpus but are generated applying our methodology and are per-

[30] <moustache$_{c_421}$ <:dobj: to_brush$_{c_932_v}$>; <mane$_{c_421}$ <:dobj: to_smooth$_{c_1267_v}$>; <fur$_{c_421}$ >:mod: silky$_{c_1223_a}$>; <goetee$_{c_421}$ >:mod: grey$_{c_149_a}$>

fectly acceptable in Spanish. These patterns would not have been extracted using, for example, a *n*-gram based method or plain statistical methods.

It is worth noting the high degree of semantic cohesion between the lemmas of the same cluster and between the lemmas of the related clusters ((8)[31], (9)[32], (10)[33] and (11)[34]).

8. <accidente $_{c\_470}$ <:dobj causar$_{c_560}$>
   <fuego $_{c\_470}$ <:dobj evitar$_{c_560}$>
   <siniestro $_{c\_470}$ <:dobj producir$_{c_560}$>

9. <accidente $_{c\_470}$ <:subj desencadenar$_{c_560}$>
   <destrozo $_{c\_470}$ <:subj producir$_{c_560}$>
   <incendio $_{c\_470}$ <:subj originar$_{c_560}$>

10. <canciller $_{c\_70}$ >:mod argentino$_{c_1}$>
    <embajador $_{c\_70}$ >:mod belga$_{c_1}$>
    <mandatario $_{c\_70}$ >:mod chileno$_{c_1}$>

11. <cantante $_{c\_155}$ >:mod belga$_{c_1}$>
    <compositor $_{c\_155}$ >:mod canadiense$_{c_1}$>
    <pianista $_{c\_155}$ >:mod estadounidense$_{c_1}$>

This strong cohesion allows for a semantic annotation of the clusters to obtain more abstract syntactico-semantic constructions that combine semantic categories (12) and (13). The semantic labels associated with each cluster have been manually added, taking into account the WordNet [Miller, 1995] upper ontologies.

12. <*Event*-n$_{c_470}$ <:dobj *Causative*-v $_{c\_560}$>
    <*Event*-n $_{c\_470}$ <:subj *Causative*-v $_{c\_560}$>

13. <*Person/Politician*-n $_{c\_70}$ >:mod *Nationality*-a $_{c\_1}$>
    <*Person/Musician*-n $_{c\_155}$ >:mod *Nationality*-a $_{c\_1}$>

[31] <accident$_{c_470}$ <:dobj to_cause$_{c_560}$>; <fire$_{c_470}$ <:dobj to_avoid$_{c_560}$>; <sinister$_{c_470}$ <:dobj to_produce$_{c_560}$>.

[32] <accident$_{c_470}$ <:subj to_trigger$_{c_560}$>, <ravage$_{c_470}$ <:subj to_produce$_{c_560}$>.

[33] <chancellor$_{c_70}$ >:mod argentinian$_{c_1}$>; <ambassador$_{c_70}$ >:mod belgian$_{c_1}$>; <representative$_{c_70}$ >:mod chilian$_{c_1}$>

[34] <singer $_{c\_155}$ >:mod belgian$_{c_1}$>; <song-writer$_{c_155}$ >:mod canadian$_{c_1}$>; <pianist$_{c_155}$ >:mod american$_{c_1}$>